Towards Accessing the Proteolytic Potential of the Gut Microbiota

Thesis presented for the Degree of Philosophiae Doctor
By

Laura Siân Morris, BSc (Hons)
Microbiology

In candidature for the Degree of Philosophiae Doctor

Organisms and the Environment Division
Cardiff School of Biosciences
Cardiff University

September 2014

DECLARATION

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed ……  …… (Candidate) Date ……8/3/2015……

STATEMENT 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

Signed……  ……Candidate) Date ……8/3/2015……

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated.
Other sources are acknowledged by explicit references.

Signed……  ……(Candidate)Date ……8/3/2015……

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed……  ……(Candidate)Date ……8/3/2015……

STATEMENT 4

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loans **after expiry of a bar on access previously approved by the Graduate Development Committee.**

Signed ……  …… (Candidate) Date ……8/3/2015……

ACKNOWLEDGEMENTS

I'd first like to thank my supervisor, Dr Julian Marchesi, who has been a constant source of knowledge, guidance, and inspiration for the past 4 years as well as some incredibly fun and memorable times. I'd also like the thank Professor Eshwar Mahenthiralingham for his advice and support throughout my PhD. I'd also like to express gratitude to the BBSRC for funding this project.

I have enjoyed every minute of studying for a PhD and this is mainly down to my colleagues, so I'd like to thank; Dr Laura Rushton, Dr Phillip Norville, Dr James Evans, Dr Matt Bull and Dr Ann Smith for their friendship, support and some very fun times. I'd particularly like to thank Miss Beky Weiser (aka "ol' greasy hands") for always being there for me, for always making me laugh, and for generally being an awesome friend/person. I know we'll be lifelong friends. I also want to thank the girls; Nash, Harris, Frost and Brown for their invaluable friendship and support for the past years and no doubt for many years to come.

I need to thank the love of my life, Lewis Osborne (aka "the Oaf") for all of his love and support, whisking me away on fun weekends to forget about my thesis, for being the most perfect boyfriend I could ever have wished for, I couldn't have done this without you and I look forward to our incredibly bright future together.

Lastly, I'd like to thank my family, and in particular my incredible, loving and wonderful parents; Stuart and Sharon Morris. Your support has been invaluable, unwavering, and overwhelming. I most certainly couldn't have got this far in my career without you. Thank you.

SCIENTIFIC AWARDS, CONFERENCES AND PUBLICATIONS

Scientific Conferences and Awards

- Society for General Microbiology spring Meeting 2012 (Dublin) Poster presentation, Received SGM travel grant.
- The 14th International Symposium on Microbial Ecology/ ISME 2012 (Copenhagen) poster presentation
- Society for General Microbiology spring meeting 2013 (Manchester)-Poster presentation,
- 5th Congress of European Microbiologists 2013 (Leipzig) Poster presentation Received SGM travel grant.
- Microbiology and Infection Translational Research Group Autumn meeting 2013 (Swansea) Received award for oral presentation.
- Rowett-INRA 2014 Gut Microbiology: From sequence to function (Aberdeen)

Peer Reviewed Publications

1. Morris, L.S., Evans, J., and Marchesi, J.R.(2012) A robust plate assay for detection of extracellular microbial protease activity in metagenomic screens and pure cultures. Journal of Microbiological Methods *91*, 144-146.

2. Evans, J.M., Morris, L.S., and Marchesi, J.R. (2013). The gut microbiome: the role of a virtual organ in the endocrinology of the host. Journal of Endocrinology *218*, R37-R47.

3. Morris, L.S and Marchesi, J.R. (2014) Current functional metagenomic approaches only expand the existing protease sequence space, but does not presently add any novelty to it. Curr Microbiol.

4. Marchesi, J .R and Morris, L. S (2014) Functional Metagenomics: Procedures and Progress. Manual of Environmental Microbiology 4th Edition. Accepeted for publication for March 2015.

SUMMARY

Introduction:-The human gut microbiota outnumbers our own human cells by 100-1 and is often considered an extension of our genome harboring fundamental functions of which we are genetically incapable. However, it can also be a significant liability and has been implicated in numerous diseases, particularly; Inflammatory Bowel Disease (IBD). The efforts of this research entailed evaluating a specific molecular mechanism of the gut bacterial metagenome; proteolytic activity, collectively known as the Degradome due to their putative role as significant virulence factors in IBD. In order to access this degradome of the gut microbiota, firstly, novel functional metagenomic (FM) tools were developed with an aim of facilitating the isolation of proteases. Secondly, a comprehensive cohort study was conducted comparing faecal protease activity, 16S rRNA microbial community structure and the potential of the faecal degradome to act as virulence factors between a group of IBD patients and a group of healthy volunteers to begin to determine the role of microbial proteases in disease aetiology.

Results:-The IBD cohort exhibited significantly higher protease activity than the healthy cohort. Inhibitor assays also showed that the IBD cohort contained different types of proteases to the healthy cohort with significantly higher levels of metalloprotease activity. 16S rRNA gene analysis of the microbial community also revealed a dysbiosis of the gut microbiota between the IBD cohort and the healthy cohort. Dysbiosis was also observed between the high protease producers and the low protease producers and protease activity in the IBD cohort was able to decrease trans epithelial resistance in an HT-29 cell line and increase cellular permeability. Functional metagenomics tools were also assessed for isolation of proteases from the gut microbiota. The ability of a protease deficient strain; *Bacillus subtilis* WB800N to express proteases was compared with *E. coli* to determine its usefulness as a host for FM screening for proteases. *B. subtilis* WB800N was able to express gelatinase E and neutral protease while *E. coli* was not suggesting *B. subtilis* WB800N was more suitable as a host. However, when the gut microbiota was screened for proteases using this host, none were isolated suggesting improvements still needed to be made. **Conclusions: -** Compositional alterations of the gut microbiota appear to be associated with high and low levels of protease activity. The IBD cohort had elevated activity and an expanded repertoire of faecal proteases which also appear to have the potential to act as a virulence factor by disrupting epithelial cell barrier integrity. Proteases remain elusive via FM screening; however this study has highlighted the areas that need improvement to optimize future screens for accessing the degradome of the gut microbiota.

LIST OF TABLES

1.1	A description of recent functional metagenomic projects	32-34
1.2	Proteases as virulence factors	41-42
2.1	Bacterial strains and plasmids used in this study	64-65
2.2	Healthy and IBD Participant information	69
4.1	The effects of metal ions and chemical reagents on protease activity	109
5.1	Proteases isolated via functional metagenomics	125
5.2	Transformation efficiencies	136
5.3	16S rRNA BLAST analysis of aerobic and anaerobic culture protease producers	139
5.4	*GelE*, *lasB* and *nprE* in *E. coli* and *B. subtilis* WB800N homology search results	139-140
5.5	Primers for amplification of *GelE*, *lasB* and *nprE*	153
6.1	Buffer composition	189
7.1	Proteases producers by pathogen in the human gut	209
7.2	16S r RNA sequencing data summary comparing IBD with healthy and Crohn's disease with Ulcerative colitis	229
7.3	Bacterial groups that are significantly different between the IBD cohort and the healthy cohort	231

LIST OF FIGURES

1.1	Publications per year in the research field of gut microbiology from 2000 to 2014	3
1.2	Intrinsic and extrinsic factors that contribute to both the shaping of the gut microbiota and altering the gut microbiota	9
1.3	A schematic diagram of the general methodology encompassing functional metagenomics and sequencing metagenomics	24
3.1	Plate assays for protease activity comparing LFSMA and SMA	86
3.2	Protease activity of *B. subtilis* MY2016 compared with 2 metagenomic clones	88
4.1	Annotation of the fosmid insert from which the two putative proteases were identified and cloned	98
4.2	Phylogenetic tree of M1-1 and M1-2	101
4.3	Sequence alignment of M1-1 and M1-2 with their closest genetic matches	102
4.4	Domain Architecture of M1-1 and M1-2	104
4.5	SDS-PAGE of cell free extracts of E. coli expressing M1-2	105
4.6	The effect of temperature and pH on M1-2 protease activity	107
5.1	Publications over the past decade when searching for functional metagenomics and sequencing metagenomics	122
5.2.1	Protease activity assay-wildtype *B. subtilis*, *B. subtilis* WB800N, and supernatant	134
5.2.2	Proteases of *E. faecalis*, *P aeruginosa* and *B. subtilis* on LFSMA	138
5.2.3	Alignments of *GelE*, *lasB* and *nprE* with their closest genetic matches	141
S 5.1	Plasmids pNZ44 digest	165
S 5.2	Partial genomic DNA digest	166
6.2.1	Boxplot of number of cycles of bead beating for optimum protein recovery	174
6.2.2	Boxplot of the effect of NaN_3 on protease activity	176
6.2.3	Boxplot of the relationship between length of storage and the decrease in protein concentrations	179
6.2.4	Boxplot of the effect of freezing temperature and the length of storage on whole faecal samples and their protein concentration.	181
6.2.5	Boxplot of the relative protease activity (%) after protein extraction of faecal samples stored for one year at -20 °C and -80°C	183

S 6.1	Protein concentration estimates and protease activity assay results over time for 3 individuals.	195
7.1	Contributory factors of development and progression of IBD	199
7.2.1	Boxplot of protein concentration estimates of the IBD cohort compared to the healthy cohort	216
7.2.2	Boxplot of azocasein, azo-coll and keratin azure assays	217
7.2.3	Temperature and pH profiles	218
7.2.4	ProteaseARREST boxplot	221
7.2.5	Protease inhibition assay boxplot	224
7.2.6	Specific protease activity profile boxplots	226
7.2.7	Mucin degradation assay	227
7.2.8	Relative abundance of bacterial phyla in IBD vs Healthy	230
7.2.9	Quartiles of protease activity	233
7.2.10	Microbial diversity of sample with high and low FP	234
7.2.11	NMDS analysis of samples exhibiting high and low protease activity	235
7.2.12	Relative abundance of bacterial phyla in high vs low protease activity samples	236
7.2.13	Effect of faecal proteases on TER of HT-29 cell line	238-239
7.2.14	Permeability assay of HT-29 cells after incubation with faecal proteases	239

LIST OF ABBREVIATIONS

ATCC	American Type Culture Collection
ATG	Autophagy related protein
BCA	Bicinchoninic acid
BLAST	Basic local alignment tool
bp	Base pairs
BGSC	Bacillus Genetic Stock Centre
BSA	Bovine Serum albumin
BNHS	Bitoin N-hydroxysuccinimide ester
CARD	Caspase recruitment domain
CD	Crohn's Disease
CDH	Cadherin
CDS	Coding sequences
cfu	colony forming units
COX	Cylcooxygenase
CRC	Colorectal cancer
DMSO	Dimethylsulfoxide
DMF	Dimethylformamide
DNA	Deoxyribonucleic acid
dNTPs	deoxyribonucleotide triphosphates
ECM	Extracellular matrix
EDTA	Ethylenediaminetetraacetic acid
EGTA	Ethylene glycol tetraacetic acid
EU	European Union
FAO	Food and Agriculture Organisation
FDA	Foods and Drugs Administration
FM	Functional metagenomics
GALT	Gut-associated lymphoid tissue
GC	Guanine-Cytosine

GIT	Gastrointestinal Tract
GNA	Guanine nucleotide-binding protein
GRAS	Generally Regarded As Safe
GWAS	Genome wide association studies
HPN4A	Hepatocyte nuclear factor 4 alpha
IBD	Inflammatory Bowel Disease
IFN	Interferon
Ig	Immunoglobulin
IL	Interleukin
IRGM	IFN-inducible protein
JAK	Janus kinase
LAMB1	Laminin Beta
LB	Luria-bertani
LFSMA	Lactose free skimmed milk agar
LRRK	Leucine rich repeat kinase
NGS	Next Generation Sequencing
MHC	Major histocompatibility complex
MRS	de Man, Rogosa and Sharpe (agar / broth)
MS	Mass spectrometry
NCBI	National Centre of Biotechnology and Information
nACHRs	Nicotinic acetylcholine receptors
NF	Nuclear factor
NOD	Nucleotide-binding oligomerisation domain-containing protein
NJ	Neighbour Joining
nMDS	Non-metric multi-dimensional scaling
nSAIDs	Non-steroidal anti-inflammatory drugs
OCP	Oral contraceptive pill
ORF	Open Reading Frame
OCTN	Organic cation transporter, novel
OPD	1,2-phenylenediamine dihydrochloride

PAI	Pseudomonas autoinducer
PAGE	Polyacrylamide gel electrophoresis
PAR	Protease activated receptor
PBD	Phosphate buffered saline
PCR	Polymerase chain reaction
PFGE	Pulse field gel electrophoresis
PI	Protease inhibitor
PTPN	T-cell protein tyrosine phosphatase
R	R statistical software
RDP	Ribosomal Database Project
RNA	ribonucleic acid
rRNA	Ribosomal ribonucleic acid
SDS	Sodium Dodecyl Sulphate
SEM	Standard Error of the Mean
SM	Sequencing metagenomics
SMA	Skimmed milk agar
SNPs	Single Nucleotide Polymorphisms
STAT	Signal transducer and activator of transcription
Taq	*Thermus aquaticus*
TCA	trichloroacetic acid
TER	Trans epithelial resistance
TLR	Toll-like receptor
TNF	Tumour necrosis factor
TYK	Tyrosine kinase
UC	Ulcerative Colitis
USA	United States of America
V	Volt
WHO	World Health Organisation

TABLE OF CONTENTS

1. INTRODUCTION 1

1.1 The Human Gut and its Microbiota **1**

1.1.1 Acquisition and development of the human gut microbiota and the core adult microbiome 3
1.1.2 Genetic and environmental factors shaping the gut microbiota 11
1.1.3 The gut microbiota as an asset 14
1.1.3.1 Metabolic function 15
1.1.3.2 Maturation of the immune system 16
1.1.3.3 Allergies 17
1.1.4 The gut microbiota and disease 18
1.1.4.1 Invasion by pathogens 18
1.1.4.2 Irritable Bowel Syndrome and Inflammatory Bowel Disease 20
1.1.4.3 Cancer 21
1.1.4.4 Obesity and related metabolic disorders 21
1.1.5 "The Unculturable Majority" and culture-independent techniques to study the gut microbiota 22

1.2 Metagenomics **23**

1.2.1 Sample preparation and DNA isolation 25
1.2.2 Sequencing-based metagenomics 27
1.2.3 Functional metagenomics 29

1.3 Microbial Proteases: The Degradome **37**

1.3.1 Microbial proteases and their uses 39
1.3.2 Proteases as virulence factors 39
1.3.3 Protease activity in the human gut 44
1.3.2 Proteases and functional metagenomics 45
1.3.3 Protease "Degradomics" 46

1.4 Aims and Objectives **47**

1.5 References **49**

2. GENERAL MATERIALS AND METHODS 63

2.1 Bacterial Strains and Plasmids **63**

2.1.1 Storage and maintenance of bacterial strains 66

2.2 Plasmids **66**

2.2.1 Plasmid extraction 66
2.2.2 Storage and maintenance of plasmids 66

2.3 Human Stool Samples **67**

2.3.1 Sample Collection 67

2.4 DNA extractions **70**

2.4.1 From bacterial cells 70

2.4.2 From faeces....71
2.4.2 Chelex rapid preparation of DNA for colony PCR....71
2.5 Restriction digests....71
2.5.1 Partial digest of genomic DNA....71
2.5.2 Restriction enzyme digestion of plasmid DNA....72
2.5.3 Polymerase chain reaction of 16S rRNA gene....73
2.6 Development of Competence in Bacterial Isolates....72
2.6.1 Development of competence in *E. coli*....72
2.6.2 Development of competence in *B. subtilis* WB800N....73
2.7 Transformation procedures....73
2.7.1 Chemical transformation of *E. coli*....73
2.7.2 Chemical transformation of *B. subtilis* WB800N....73
2.7.3 Electrotransformation procedure....74
2.8 Activity based screening....74
2.8.1 Preparation of semi skimmed milk agar (SMA) and Lactose-free skimmed milk agar (LFSMA)....74
2.8.2 Preparation of crude cell free extracts....75
2.9 Measuring protease activity....75
2.9.1 Azocasein assay....75
2.9.2 Azo-coll assay....76
2.9.3 Keratin azure assay....76
2.10 Protein Extraction....76
2.10.1 Total and extracellular protein extract from faecal samples....76
2.11 SDS PAGE....77
2.12 Statistical Analysis....77
2.13 Cell Culture....77
2.13.1 Handling of frozen cell suspension and preparation of initial low density cell stock....78
2.13.2 Cell line maintenance and sub-culturing procedure....78
2.13.3 Cryopreservation....78
2.14 References....79

3. A ROBUST PLATE ASSAY FOR DETECTION OF EXTRACELLULAR MICROBIAL PROTEASE ACTIVITY IN METAGENOMIC SCREENS AND PURE CULTURES....80
3.1 Introduction....80
Aims....84
3.2 Results....85
3.2.1 Agar plate assay....85
3.2.2 Azocasein assay....87
3.3 Discussion…. ….89
3.4 Material and methods....90
3.4.1 Strains and culture conditions....90

3.4.2 Preparation of media 90
3.4.3 Azocasein assay 90
3.4.4 Statistical analysis 90
3.5 References…. 91

4. SREENING, ISOLATION AND CHARACTERISATION OF TWO FUNCTIONAL METAGENOME-DERIVED METALLOPROTEASES ISOLATED FROM A MILK WASTE TREATMENT PLANT 93
4.1 Introduction 93
Aims 96
4.2 Results 97
4.2.1 Sequence analysis of M1-1 and M1-2 putative protease-encoding sequences 97
4.2.2 Activity based screening of sub clones 106
4.2.3 Characterization of M-12 protease 106
4.3 Discussion…. 110
4.4 Material and methods 114
4.4.1 Strains and plasmids and growth conditions 114
4.4.2 Sub cloning and activity based screening 115
4.4.3 Preparation of cell free extract and enzyme characterization 115
4.5 References…. 117

5. DEVELOPMENT OF FUNCTIONAL METAGENOMIC TOOLS AND SCREENING THE GUT MICORBIOME FOR PROTEASE ACTIVITY 120
5.1 Introduction 120
Aims 129
5.2 Results 131
5.2.1 *B. subtilis* WB800N adaptation for functional metagenomic screening for proteases 131
5.2.2 Culture-dependent identification of protease producing faecal bacteria 135
5.2.3 Comparison of expression of characterized proteases in a Gram negative and a Gram positive host 137
5.2.4 Use of a Gram negative- Gram positive host shuttle vector system to screen the human gut for proteases 142
5.3 Discussion…. 143
5.4 Material and methods 151
5.4.1 Culture-dependent analysis of protease-producing bacteria in the gut microbiota 151
5.4.2 Amplification of *gelE*, *nprE* and *lasB* by PCR 153
5.4.3 Cloning and screening of *gelE*, *nprE* and *lasB* into pNZ44 154
5.4.4 Metagenomic library construction and screening 155
5.5 References…. 156
5.6 Supplementary material…. 163
5.6.1 ‘Paris’ medium protocol for *B. subtilis* WB800N transformation 163

5.6.2 Spizizens protocol for *B. subtilis* WB800N transformation 163
5.6.3 MD Medium protocol for *B. subtilis* WB800N transformation 164
5.6.4 Electrotransformation protocol of *B. subtilis* WB800N 164
5.6.5 Plasmid (pNZ44) digest 164
5.6.6 Genomic DNA partial digest 166

6. THE EFFECT OF STORAGE CONDITIONS ON THE REPRUCIBILITY OF PROTEIN CONCENTRATION AND PROTEASES FOR ANALYSIS OF THE FAECAL MICROBIOTA 167
6.1 Introduction 167
Aims 173
6.2 Results 174
6.2.1 Effect of number of bead beatings on protease activity 174
6.2.2 Effect of NaN_3 on protease activity 175
6.2.3 Stability of protein concentration and protease activity of faecal protein extract over time 177
6.2.4 The effect of freezing temperature of whole faecal samples and reproducibility of protein concentration and protease activity 180
6.3 Discussion…. 185
6.4 Material and methods 189
6.4.1 Buffer composition and preparation. 189
6.4.2 Collection of faecal samples 189
6.4.3 Protein extraction 189
6.4.4 Protein concentration measurements and protease assays 189
6.4.5 Design, conduct and analysis of storage media for evaluation of protein yield and protease activity over time 191
6.4.6 Design, conduct and analysis of the effects of storage of whole faecal samples at -20 and -80 °C 191
6.4.7 Statistical analysis 191
6.5 References…. 192
6.6 Supplementary material…. 195
Protein concentration estimates and protease activity assay results over time for 3 individuals. 195

7. A COMPREHENSIVE COMPARISON OF THE HEALTHY IBD FAECAL DEGRADOME 197
7.1 Introduction 197
Aims 215
7.2 Results 216
7.2.1 Protein concentration in healthy and IBD protein extracts 216
7.2.2 Faecal Protease activity in Patients with IBD compared to Healthy controls 217
7.2.3 Temperature and Ph profiles of FP activity 219

7.2.4 Protease inhibitor assays 226
7.2.5 Specific protease activity profiles 227
7.2.6 Degradation of mucin 228
7.2.7 Bacterial Community in the Healthy cohort and IBD patient cohort 232
7.2.8 Association of Faecal Microbial Diversity and Taxonomy with protease activity 237
7.2.9 Effect of protease activity on TER 238
7.3 Discussion…. ..... …. 241
7.4 Material and methods 258
7.4.1 Characterization of faecal proteases within the healthy cohort and the IBD patients cohort 258
7.4.2 Protease inhibitor assays 259
7.4.3 Optiprep™ density gradient for isolation of bacterial cells 259
7.4.4 Protease assays using specific substrates 260
7.4.5 Biotinylation of mucin 260
7.4.6 MiSeq™ Illumina sequencing of 16S rRNA genes 262
7.4.7 Analysis of 16S rRNA gene sequences using the Mothur Pipeline 262
7.4.8 Transepithelial resistance measurements 263
7.4.9 Permeability assay 263
7.5 References…. ..... …. 264

8. GENERAL CONCLUSIONS AND DISCUSSION 277
8.1 Conclusions 278
8.2 Concluding remarks 284
8.3 References…. ..... …. 285

1.0 INTRODUCTION

1.1 The Human Gut and it's Microbiota

The human gastrointestinal tract is the site where the food we intake undergoes its metabolic journey, its catabolised as much as possible, and converted to the essential energy, nutrients and other products of metabolism that are necessary for life. The human gut has its own nervous system (enteric or intrinsic nervous system) that is capable of functioning autonomously, carrying out numerous physiological and pathophysiological processes including; modulating exocrine and endocrine functions, gastrointestinal blood flow, motility and modulation of enteric immune functions and consequently epithelial homeostasis [1]. Clearly, the human gut is a vastly complex organ. Aside from these fundamental functions, this nutrient-rich environment is also the 'home' to trillions of microbial inhabitants that exist in numbers exceeding those found anywhere else in the human body, and that have co-evolved to form an intrinsic, interdependent relationship with us, as their hosts, to a point where the absence of our gut colonisers could have significant effects on our health. This intricate and delicate relationship has led, unsurprisingly, to the human gut microbiota becoming a major area of research in the scientific world with a major focus on determining it's abundance, diversity, function and how we influence them but equally how they influence us.

The gut microbiota is comprised of viruses, eukaryotic microorganisms and *Archaea*, but these are outnumbered both in cell density and phylogenetic diversity by the kingdom: *Bacteria*. This knowledge, and our ever increasing knowledge of the gut microbiota has been significantly aided by the advent of culture-free molecular based techniques, particularly bacterial community profiling by 16S rRNA gene sequencing analysis and shotgun metagenomics. Both such strategies by-pass the need to culture, as we are now aware that culture techniques available now are currently only able to support the growth of as few as 1% of total microbial diversity (for more information please Chapters 3 and 5). The 16S rRNA gene is commonly used to carry out diversity analysis of bacteria due to its uniform presence amongst bacteria and the ability to identify species based on unique domains. Metagenomics has become a term used to encompass all the genes in their entirety that are isolated from a particular environment. While there are sections in this thesis dedicated to the detailed description of

metagenomics (section 1.3) in brief, the process is characterised by; the isolation of DNA from an environment of interest, random shearing of the DNA using various methods and reconstruction of sheared fragments into vectors to create a metagenomic library. This library is subsequently sequenced by way of the increasingly high-throughput DNA sequencing methods. The cloning step can also be side-stepped and samples can be subject to next generation sequencing straight away. These include technologies such as Roche/454 pyrosequencing, Illumina Genome analyzer and the SOLiD™ Sequencer (more details of which will be discussed further on in section 1.3). Generated data can be vast (gigabases of nucelotide sequencing data) and must be made sense of using advanced bioinformatics using a variety of computational tools.

The characterisation of microbial communities using 16S rRNA gene sequences and metagenomics in pinnacle projects that have completely altered the way we view microbial ecology conducted on human faecal samples (as well as other parts of the human body) such as those conducted by Eckburg *et al*.,[2] large scale international projects such as the United States National Institute of Health sponsored Human Microbiome Project (HMP) [3] , the Metagenomics of the Human Intestinal Tract (MetaHIT) project [4, 5] and ELDERMET [6] (http://commonfund.nih.gov/hmp/) have helped uncover the magnitude of microbial abundance, diversity and functional potential. These data suggest that there are over 1000 different bacterial species can inhabit the human gut [7] and a possible 8000 bacterial strains [8] and an individual can harbour around 160 different bacterial species at one time [4]With data such as this it has been estimated that the human gut microbiota harbours ≥ 100 times more genes than the human genome of ~3 billion base pairs [8] (Ley, 2006). Understanding diversity and abundance of the gut microbiota has also been key in providing the stepping stones for going on to determine what these 300 or so billion base pairs of DNA are genetically capable of and the implications for us as their hosts. For example, we now understand that we have coevolved a mutually beneficial relationship with our gut microbiota, and that their genetic repertoire confers us with necessary metabolic attributes that we are not capable of such as the digestion of otherwise indigestible protein and polysaccharides of which the metabolites can be utilised by our cells [9] . It has also been demonstrated that perturbations in the gut microbiota can contribute to a plethora of diseases. It is unsurprising therefore, that there is significant and increasing

interesting research occurring in the field of gut microbiology as demonstrated by figure 1.1

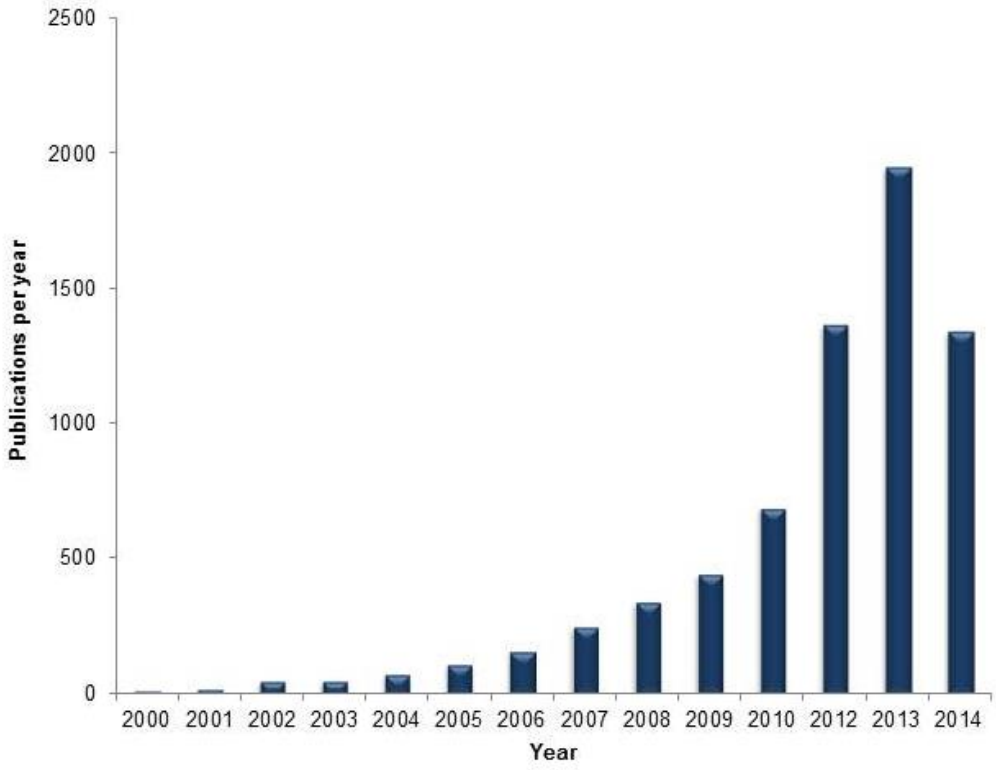


Figure 1.1. Publications per year in the research field of gut microbiology from 2000 to 2014. Adapted from [10] Figure 1 helps draw attention to just how much interest is being gathered in the area of gut microbiology since the increased implementation of culture-independent techniques and next generation sequencing (NGS) technologies showing the increase in publications since 2000 to present. Values for the figure were obtained using the keywords; 'gut microbiota' from the ISI Web of knowledge database (http://apps.isiknowledge.com/).

1.1.1 Acquisition and development of the human gut microbiota and the Core Adult Microbiome

Microbial colonisation begins immediately after birth as in the womb, infants are widely believed to be born sterile [11] however recent research such as those conducted by Jiménez and colleagues [12, 13] have challenged this concept having isolated bacteria from umbilical cord blood and amniotic fluid of neonates [13] as well as from the meconium of murine neonates [12]. Acquisition of the gut microbiota is incredibly

complex and driven by both external and internal factors. Research has shown that it is highly dynamic, subject to dramatic change and relatively low in diversity compared to bacterial communities of the healthy adult gut that reaches numbers previously described.

A pioneering study into the acquisition of the gut microbiota was undertaken by Ley *et al.*, comparing gut bacterial community compositions of related mice [9]. They found that mothers and offspring shared similar community profiles as did siblings and different litters from the same mother, but litters from different mothers had significantly different bacterial communities thus suggesting that microbial communities are inherited from mothers. The bacterial communities are 'inherited' predominantly from mothers via vertical transmission i.e. skin-oral transmission and faecal-oral transmission [14] and also the external environment upon birth and human genotype also indirectly contribute to determining an individual's gut microbiota by determining their immunological phenotype [15]. Culture-dependent and independent analysis have deduced that 'pioneer' gut bacteria to predominantly include members of the *Enterobacteria*, *Bifidobacteria*, *Eneterococci*, *Bacteroides* spp, *Clostridia* and *Ruminococcus* spp.[14]. However, there are a number of factors that can also influence the first bacterial colonisers of a new-born; research has shown that initial bacterial species differs as a result of mode of delivery and mode of feeding [14]. For example, vaginal delivery results in a microbiota resembling that of the vaginal microbiome with *Lactobacilli* dominating and lower numbers of the *Enterobacteriaceae* family and *Enterococci* phyla compared with Caesarean-section delivery where bacteria that are usually found inhabiting skin such, i.e. *Staphyloccous* spp. and *Propionobacteriums* spp. are more likely to be the dominant species [16]. Though their poor adaptation to the human gut leads to significant decline in following months [17]. Other factors affecting this early colonisation have also been studied, such as feeding, prematurity and exposure to antibiotics. Exclusively breast-fed infants' gut bacterial communities differ from exclusively formula-fed infants according to research by Penders *et al*., [18] Significant species found colonising the formula-fed infants in significantly higher numbers include *E. coli, Clostridium difficile, Lactobacilli* and *Bacteroides fragilis* [18]. The use of antibiotics and hygiene as well as other lifestyle related factors have also been shown to influence the infant gut microbiota [17, 18].

Following this preliminary colonisation of these 'pioneer' bacteria [15], the gut microbiota develops in complexity and diversity throughout the first year of life, although it has been demonstrated that the microbial communities remain distinct [19] after such time the gut microbiota begins to converge to resemble that of an adult in that species commonly found in infants, but rarely in adults begin to decline and disappear and species commonly found in adults begin to dominate. However, immune, genetic and other environmental factors also play a role in determining an individuals' gut microbiota (to be discussed later). The adult gut microbiota is also considerably more stable in that the community profile of an individual will remain more comparable over time than it will to another individual's gut microbiota. An adult human has been estimated to harbour as many as 100 trillion bacterial cells, encoding at least 100 fold more genes than there are on our own genomes [8] and the majority of which are found in the gut [2]. Most of the pioneering studies analysing the gut microbiota implementing culture-independent techniques (that include temperature gradient gel electrophoresis analysis of 16S rRNA genes [20], denaturing gradient electrophoresis analysis of 16S rRNA genes [21] and fluorescence *in situ* hybridisation (FISH) [22] and more recent analyses have implemented high throughput culture independent techniques; direct 16S rRNA gene sequencing coupled with community profiling and metagenomics [2, 23] on samples from healthy adult volunteers. The definative work of Eckburg and colleagues [2] analysed mucosal tissue samples from the colon, cecum, ascending colon, transverse colon, descending colon, sigmoid colon and rectum by colonoscopy from three healthy individuals as well and faecal samples from the three individuals taken 1 month after colonoscopy. In total 13,335 16S rRNA gene sequences were analysed and over 395 bacterial phylotypes were identified and, although 1524 archaeal sequences were identified, it was found that they all belonged to just one archaeal phylotype; *Methanobrevibacter smithii* highlighting the difference in diversity between the two prokaryotic Kingdoms. The vast majority of bacterial sequences analysed contained novel species (62%) and 80% sequences were from species that were yet to be cultivated thus highlighting the extent of novelty in the gut. This study also revealed most sequences isolated from the gut belonged to the *Firmicutes* (most of which were from the *Clostridiales*) and the *Bacteroidetes* phyla and fewer numbers belonged to *Proteobacteria, Actinobacteria, Fusobacteria* and *Verrucomicrobia* [2]. These data were concordant with previous findings that had implemented 16S rRNA gene analysis techniques towards characterising the gut microbiota [24, 25] as well as 16S rRNA gene

analysis coupled with anaerobic culture methods [26]. These studies have also shown that there is distinct variation between individuals, but also intra-sample variations along different anatomical sites of the gut [2].

Fluctuations in the gut microbiota have been shown to have profound impact on host health. Understanding the diversity and genetic and functional potential of the gut microbiota is fundamental to understanding the role they have in human health and this is why recent years have seen increased effort in trying to decipher the gut microbiome and firstly elucidating the presence of microorganisms that might be indicative of a 'healthy' state [27]. While studies on the gut microbiota still often report variations within and between individuals, it has still been put forward that, if we are to begin to understand the precise roles the gut microbiota has on health and disease we need to establish whether or not there is a 'core' microbiome found within at least a majority of individuals regardless of external factors (diet, ethnic origin, lifestyle etc.) that may also influence the gut microbiota. To begin to answer this question, Arumugam *et al.*, conducted an interesting study to determine if there were microbial commonalities between different global populations. They conducted a large cohort based study spanning samples from Danish, French, Italian and Spanish individuals and compared the sequenced metagenomes of these individuals with existing databases comprising Japanese and American samples. One of this study's key finds was that individuals from the entire cohort clustered into 3 distinct clusters which were enriched in certain bacterial species. These clusters became known as enterotypes (which could be thought of a much less stringent human blood type sort of system of characterising populations) which clustered independently of continent or nation. The three enteroptypes were groups that were enriched in; *Bacteroides*, *Prevotella* and *Ruminococcus* and additionally functional profiles exhibited a similar clustering suggesting function of a particular population could be estimated depending on their enterotype [28].

Qin *et al.*, [4] conducted a study as part of the MetaHIT consortium to both determine core species and core functions of the gut microbiota. In this study, 124 volunteers were recruited, including healthy individuals and individuals with obesity and IBD. While bacterial communities were found to cluster depending on disease state, they also found 18 species that spanned the entire cohort at a 1% coverage rate. Core species included *Faecalibacterium prausnitzii*, *Roseburia intestinalis* and *Bacteroides uniformis* the

beneficial attributes of which will be discussed further (section 1.13.1). The study sought to determine functions of the gut microbiota that might be indicative of adaptation to life in the gut. They hypothesised core functions should appear in all or at least most of the bacterial species present at a frequency higher than that of functions found only in a limited number of species. Main function uncovered following metagenomic analysis revealed housekeeping functions that are similar for most bacteria regardless of habitat and include functions such as amino acid synthesis and carbohydrate metabolism. However, putative gut specific functions were also put forward and these including functions that appeared to be involved in bacterial adhesion to host protein such as collagen and fibrinogen. However over 80% of the putative orthologous group uncovered contained genes that are currently of unknown function indicating that much work in this area is still to be done if we are to uncover the gut function of individuals or populations. Definite functions uncovered included a number of genes involved in the degradation of complex sugars and glycans indicating that dietary sugars must be a core energy source of the gut microbiota [4]. Similarly, a large scale metagenomic study by Gill *et al.,* found the gut microbial metagenome to be enriched in genes involved in the degradative pathways of plant polysaccharide, glycan and amino acid synthesis [23]. An interesting study conducted by Hehemann and colleagues [29] has also demonstrated how gut bacteria have acquired such functional genes. In their study, they identified genes encoding porphyranases first characterised from a marine *Bacteroidetes* species; *Zobiella galactanivorans* in the gut bacterium *Bacteroides plebius* in a Japanese population. Zobiella galactanivorans is an inhabitant of the seaweed that makes up a proportion of the Japanese typical diet. These data suggest a mechanism for how polysaccharide degrtading enzymes, such as porphyranases, were acquired by the gut microbiota and also how the function of the gut microbiota may be governed by the host diet [29].

Larger studies are now needed to determine whether these findings are synchronous, and to also begin to investigate how core functions of the gut microbiota may be contributing the state of the host (i.e. healthy/ diseased). Lozupone *et al.,* suggest that the concept of a core microbiome is becoming more implausible as data sets expand globally [30]. While some studies discussed here demonstrate a level of concordance between a core microbiome globally, other studies have found contrasting results [31-33]. The contribution of metagenomics and invaluable use of animal, mainly mouse

models have shown that the gut is enriched for functions that significantly contribute to host health and gut function. An arsenal of tools are now available and will be increasingly progressed in order to unravel the fundamental functions that contribute to host health and gut function.

The environment of the gut puts enormous selection pressure on the bacterial community [8] for example, the ability to adhere to the mucus lined gut epithelia cell lining, a repertoire of degradative enzymes to metabolise the vast array of nutrients available and out-compete other microorganisms, the ability to evade or avoid the host immune system and the ability to grow fast and adapt should it be required, are all obligatory features of which a microorganism must be genetically capable in order to establish and maintain a community in the human gut. Bacterial density and diversity fluctuate throughout the length of the gut. Figure 1.2 highlights some of the key intrinsic factors that selects the bacterial species present and affects distribution of bacteria including the molecular mechanisms that allow them to inhabit in some cases particularly harsh environments.

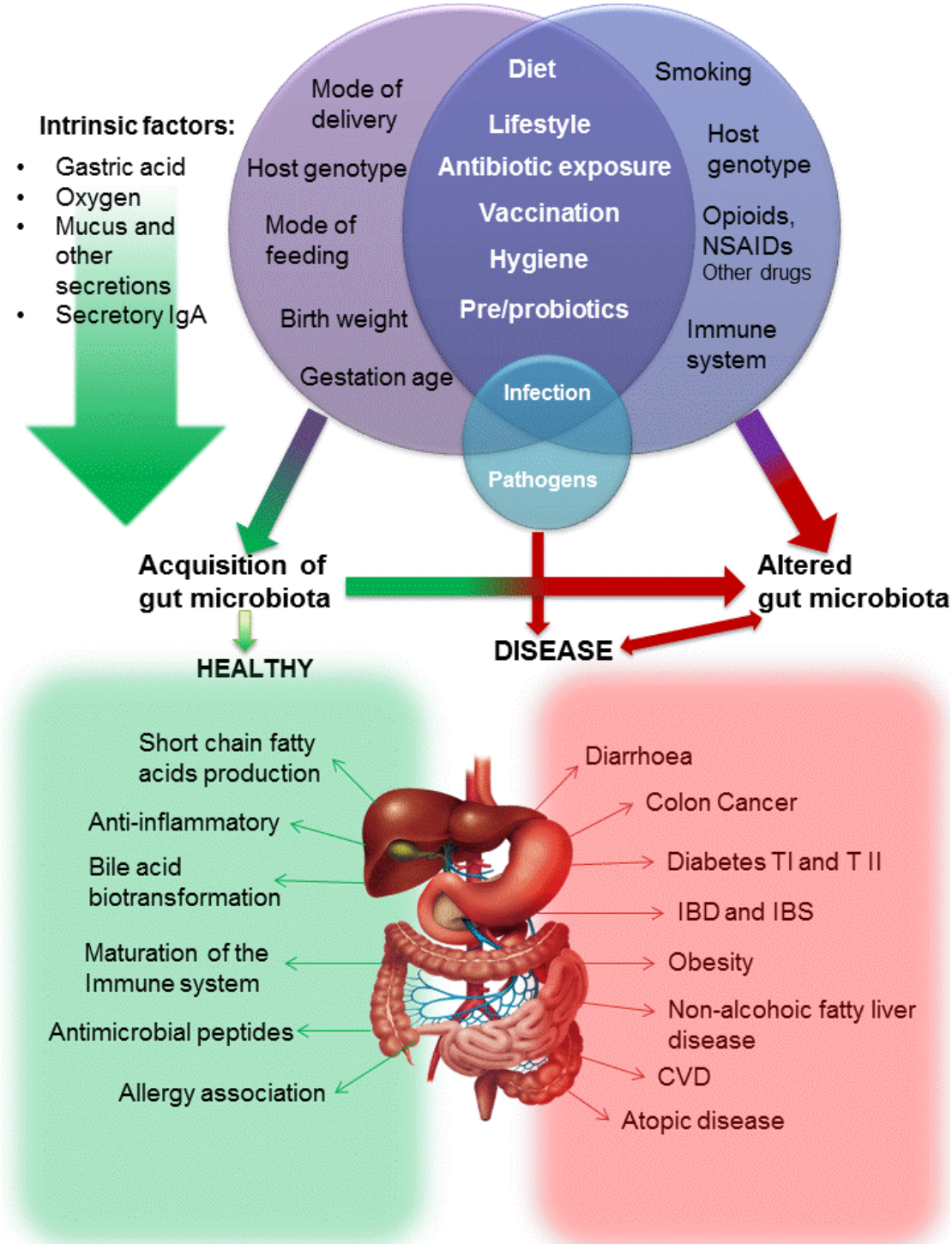


Figure 1.2. Intrinsic and extrinsic (in the purple circles) factors that contribute to both the shaping of the gut microbiota and also in altering the gut microbiota and how the gut microbiota affects health and disease. Intrinsic factors of the host are often different at different anatomical sites of the gut and govern the bacteria that can inhabit. Some extrinsic factors have an effect on both acquisition and alteration of the gut microbiota as shown by the factors in the overlap of the circles. A 'healthy' gut

←*Continued from previous page*

microbiota lives harmoniously with its host conducting essential metabolic and protective functions. It is currently unknown whether the altered gut microbiota (dysbiosis) is a cause or effect of certain diseases such as those shown here. Either way, dysbiosis of the gut microbiota is associated with a plethora of disorders both specific to the gut and elsewhere in the human body. Extrinsic factors may very well be directly altering a gut microbiota to favour those capable of contributing to such diseases.

Wang and colleagues compared the bacterial diversity from biopsies of the mucosa of the human jejunum, distal ileum, ascending colon and rectum using 16S rRNA gene analysis and found that the jejunum was the least diverse and was dominated by species belonging to the *Streptococcus* genus (Firmicutes phyla). Diversity increased in the ileal and colonic samples with the *Bacteroides* phyla and *Clostridium* clusters XIVa and IV (part of the Firmicutes phyla) dominating [34] which was concordant with many subsequent analyses previously mentioned [2, 23] which were conducted using advanced next generation DNA sequencing technologies.

Bik *et al.* also used 16S rRNA gene sequencing for community analysis of the human stomach, which had been previously concluded to harbour very few bacterial genera; *Helicobacter* [35] and *Lactobacillus* [36] due to their ability to survive the acidic gastric pH as a result of urease production [37] and regulation of a extracellular and cytoplasmic pH gradient [38]. 16S rRNA analysis revealed a more abundant and diverse bacterial gastric community than previously thought with many species present previously undescribed by culture techniques including the genera *Caulobacter, Actinobacillus, Corynebacterium* and *Deinococcus* [39], thus highlighting both the strength on sequencing and the ability of some bacterial genera to withstand such extreme environments.

When Bacterial density reaches its peak, they outnumber our own cells by at least 100-1 and subsequently also outnumber our own genes. As previously alluded, there is significant urgency to determine what these 300 or so billion base pairs of DNA are genetically capable of and how they impact us, as their hosts. [8]. The collective genes of all the microorganisms within the human gut (the microbiome) can be thought of as

an extension of our genome conferring genetic attributes that we are not capable of. Bacteria have been found to have such a distinctive role within the human gut that now many microbiologists may often refer to it as 'the forgotten organ'.[40]. Is the presence of any particular organisms indicative of health? To answer this question it is first essential to understand their functional and genetic repertoire.

Aims put forward by many large scale projects (for example, The human microbiome project, MetaHIT) are to include a wider data set of individuals from heterogeneous backgrounds that have had their gut microbial community sequenced and identified to try to help us answer the fundamental question; is there a core microbiome associated with a 'healthy' gut? A question that can only be answered with a much expanded data set and analysis of the functions the gut microbiota has specifically regarding the effect they have on the host.[7]. It is also paramount to understand the external factors that contribute to shaping and altering the gut microbiota that are likely having a profound effect on gut microbiota function.

1.1.2 Genetic and Environmental factors shaping the gut micrbiota

While the human gut microbiota is variable from one individual to the next, studies have shown that family members tend to harbour more similar microbial communities [9, 41]. Theories emerged that similar lifestyles and similar environmental factors such as diet and hygiene (figure 1.2) have a profound impact on the gut microbiota and these will be discussed in more detail further on in this section, however it has been demonstrated that host genotype plays a substantial role in shaping the gut microbiota in terms of bacterial communities [8] conversely, this has been demonstrated to not be the case for viral communities [42]. Studies focused on the 'heritability' of the gut microbiota through twin studies, mouse lines and quantitative trait loci (QTL) have provided valuable insights into how host genotype contributes to the development of the gut microbiota. Heritability of the gut microbiota refers to the genetic basis of development of the gut microbiota as opposed to non-genetic acquisition of the gut microbiota previously discussed.

Benson and colleagues utilised mouse models and QTL analysis to determine whether specific bacterial taxa co-segregated as quantitative traits linked with genomic

polymorphisms. Using 520 polymorphism markers, 18 host QTL were identified that showed significant association with relative abundance of specific bacterial taxa. It was also found that QTL had 3 means of affecting microbial abundance by loci either controlling one particular species, groups of related species or one loci could influence multiple, and distantly related bacterial taxa [43]. However, studies have also revealed equally similar gut microbiota in monozygotic twins as well as dizygotic twins strongly suggesting a role for environmental and not just genetic factors in shaping the gut microbiota [41].

Host diet is a significant contributory factor and has been comprehensively reviewed [44-46]. Changes in diet or differences in diet due to geographical location and lifestyles cause a difference in the substrate available within the large intestine as exemplified by the previously discussed work of Hehemann for example [29]. Different substrates may also initiate changes in the gut environment in terms of pH, oxygen and gut transit time also of which can significantly contribute to determining the taxonomy of bacterial species present on the gut due to their different genetic composition and metabolic capacity [44]. Studies investigating the effect of specific diets characteristic of different parts of the world have highlighted how diet causes changes in abundance of microbial taxa and also changes in metabolic pathways. Perhaps some of the more prestigious work on diet-induced changes of the gut microbiota includes the work of Turnbaugh and colleagues [47] who presented a study where human adult faecal samples were transplanted into germ free C57BL/6J mice and implemented metagenomic analysis to monitor the heritability of the microbiota. It was found that the mice became colonised with a microbiota comparable to the donor. These mice were subject to dietary alterations, from a low fat, plant based, polysaccharide rich diet to a high fat, high sugar diet typical of the Western world. After only, 1 day, shifts in their microbiota were observed in terms of not only diversity, but also metabolic pathways and gene expression. It was also found that mice fed on the western diet had a significant increase in adiposity and that this trait could be transferred to other mice via faecal transplant [45] strongly inferring that the gut microbiota can contribute to an increase in fat storage.

Diets high in protein and animal fat had a characteristic association with members of the *Bacteroides* and diets high in carbohydrate were strongly associated with *Prevotella*

[48]. In the second part of their study, 10 volunteers were recruited and subject to a controlled feeding study where one group were fed diets high in fat and low in fibre and the other group were fed the converse. While dietary changes were shown to alter the entire gut microbiome, enterotypes remained stable, thus suggesting that short term dietary changes do not alter gut enterotypes, but that these changes are more dependent on long term dietary habits [48].

Clearly, diet can alter specific taxa of the gut microbiota and studies such as those by Turnbaugh *et al.,* [47]have shown the metabolic consequences. Thus diet has been increasingly implicated in disease related to the gut microbiota, particularly obesity which will be discussed in more detail later on. David and colleagues [46] recently set about to determine how dietary interventions alter the gut microbiome and found that short term animal-based diets increased microorganisms such as *Alistipes, Bilophila* and *Bacteroides*; these are bile tolerant organisms, and the levels of *Firmicutes* were found to be reduced. Conversely, the diet composed of mainly plant polysaccharides constituted an increase in *Roseburia, Eubacterium rectale* and *Ruminococcus bromii*, organisms that have a demonstrated ability to metabolize plant polysaccharides [46]. Clearly, microbial metabolic capacity is a key driving force behind changes in the gut microbiota in response to dietary changes.

Lifestyle choices other than diet such as sanitation, urbanisation, use of antibiotics and vaccinations have also been demonstrated to have various levels of impact on the gut microbiota and consequently, alter intestinal metabolism having potential implications on human health. The increasing use of broad spectrum antibiotics has received attention for numerous factors, one of which is the effect antibiotic usage are having on perturbations on the gut microbiota. It is widely accepted that antibiotics have profound short term effects on the gut microbiota as demonstrated by studies such as that of Pérez-Cobas and colleagues [49] who implemented the use of multiple ‘omic’ based strategies to monitor how antibiotics caused disturbances of the gut microbiota. Unsurprisingly, a significant drop in diversity was initially observed along with a decrease in protein expression determined by metaproteoimcs and an enhanced expression of certain proteins involved in glycolysis, pyruvate decarboxylation the tricarboxylic acid cycle, glutamate metabolism, iron uptake, GTP hydrolysis and translation termination but protein expression alterations reverted back to more normal

levels after treatment was completed [49]. Some studies have also sought to determine the more long term effects antibiotic usage may be having on the gut microbiota. Jernberg *et al.,* [50] studied the effect of a 7-day treatment of clindamycin on the gut microbiota, but more specifically, *Bacteroides*, for a period of 2 years using repetitive sequence-based PCR and real-time PCR. It was found that *Bacteroides* isolates had a marked decrease in clonal diversity after antibiotic usage and a rapid enrichment of resistant clones was also observed. The overall level of resistance genes after antibiotic usage significantly increased and *Bacteroides* fluctuated throughout the 2 year sampling period and did not return to its original composition suggesting long term change in the gut microbiota due to antibiotic exposure [50].

Perhaps related to antibiotic treatment, invasion by 'new' species may also have consequence on the diversity of the gut microbiota. Previous data have shown that the resident microbiota contribute to preventing invasion by potentially harmful bacteria [51, 52] and generally forms the basis behind the development of many probiotics. Manichanh *et al*., [53] however challenged this quite well established concept and determined that the gut microbiota could be much more easily manipulated than previously anticipated [53]. Perhaps the most widely known example of a coupling of antibiotic therapy and invasion of a new species in altering the composition of the gut microbiota is *Clostridium difficile* associated diarrhoea. Here, normally a hospitalised person who is receiving antibiotics for various reasons, becomes infected with *C. difficile* as a result of a perturbed and reduced gut microbiota. Symptoms include severe diarrhoea, abdominal pain and fever. The success of faecal transplantation in ameliorating this disease has highlighted the plasticity of the gut microbiota [54]. This approach also highlights some of the beneficial attributes a gut microbiota can confer.

1.1.3 The Gut Microbiota as an Asset

The relationship between the mucosal surface of the gut epithelia and it's microbiota is one of mutual symbiosis. Here the trillions of microorganisms can conduct a repertoire of metabolic, structural and defensive functions that are beneficial to their human host and in return, they have a nutrient rich environment in which they can flourish.

1.1.3.1 Metabolic function

Gut bacteria are responsible for many metabolic reactions that lead to the subsequent availability of valuable nutrients and energy such as amino acids, vitamins including; vitamin B, riboflavin, pantoic acid vitamin $B_{12,}$ vitamin K, nicotinic acid, thiamine and biotin [55, 56] and are fundamental in the metabolism of certain, otherwise indigestible complex carbohydrates, proteins, dietary fibre or plant polysaccharides and fat predominantly by *Bacteroides* and *Bifidobacterium* species [57, 58] (figure 1.2 highlights some of the benefits of a gut microbiota). Metabolism of complex carbohydrates by the gut microbiota provides the host with a variety of end products, the most notable of which relating to gut health is that of short chain fatty acids (SCFAs). SCFAs include acetate, butyrate and propionate and are involved in modulating inflammation, wound healing and motility in the gut and also act as substrates for energy conversion For example; acetate has been shown to increase cholesterol synthesis, butyrate acts as main energy source for colonocytes [58, 59] and propionate acts as the main energy source for peripheral tissue [60] [59] [61]. Some bacteria that have been identified in a majority of healthy individuals from sequencing projects have been further studied to determine their metabolic function and the consequence of this on host healthy with regards to determining whether putative 'core' species are beneficial to their hosts. For example, *Roseburia intestinalas* has been demonstrated as a butyrate producing organism [62] as have other species such as *Butyricoccus pullicaecorum* [63]. Numbers of *B. pullicaecorum* as well as *Faecalibacterium praznitzii* have been demonstrated to often significantly decline in the gut microbiota of IBD patients [64] instigating research into their role as mediators of inflammation in the gut. Sokol and colleagues [65] assessed the anti-inflammatory properties of *F. prausnitzii in vitro* and *in vivo* and found that *F. prausnitzii* supernatant could reduce interleukin-8 (IL-8)-a pro-inflammatory cytokine, secretion by a human gut epithelial cell line (Caco-2) as well as eliminate nuclear factor-kappa B, a protein that controls DNA transcription of genes such as cytokines, chemokines and as adhesion molecules as part of the pro-inflammatory response [66]. The *F. prausnitzii* also attenuated the effect of colitis in mouse models of inflammation and increased the secretion of IL-10, an anti-inflammatory cytokine, and reduced secretion of Tumour necrosis factor alpha (TNF-α) and Il-12 which are both pro-inflammatory molecules [65]. Butyrate production, another characteristic of *F. prausnitzii*, as well as other putative beneficial bacteria previously

mentioned, can also exert anti-inflammatory effects through inhibition of NF-$_K$B [67]. Clearly there is evidence to suggest a beneficial role of the gut microbiota perhaps even down to the species level.

The gut microbiota also plays a profound role in the biotransformation of bile acids. Bile acids are synthesised from endogenous cholesterol in the liver and consists of a steroid core conjugated to glycine or taurine which is secreted into the intestine where they carry out their main role of solubilising dietary fat. While this is a fundamental role, the detergent –like nature of these molecules means they can be toxic to colonocytes and have been linked with the pathogenesis of colon cancer [68]. The ability to metabolise bile acids to non-toxic metabolites (and sometimes toxic metabolites such as deoxycholic acid which is notably a by-product of microbial metabolism)is an attribute found in many commonly isolated species of the gut microbiota including *Bacteroides, Clostridium* and *Bifidobactera* [69]. The bacterial enzymes responsible are known as bile salt hydrolases (BSH) and have been found to be a conserved function within the gut microbiota [70] and their activity on bile acids results in free primary bile acid (deconjugated) and amino acids. BSH enzymes therefore detoxify bile acids and are often a characteristic function of bacteria used in probiotics [71]. However use of BSH activity probiotics remains a controversial subject as BSH activity has also been noted to exert detrimental effects on the host [71] and because metabolism sometimes yield toxic by-products.

1.1.3.2 Maturation of the Immune system

Perhaps one of the more defined positive activities of the gut microbiota is it's role in the maturation of both the innate and adaptive immune response. Early exposure to commensal bacteria in the gut is involved with immune system adaptation and tolerance to such bacteria. Germ-free animals compared with their normally inhabited counterparts have demonstrated that animals kept under sterile conditions have reduced circulatory levels of specific and non-specific antimicrobial components such as Immunoglobin A which are essential for protecting the mucosal barrier [72].

In addition to their protective role in developing the immune system, pathogens also have to compete for survival with these long established commensals making it much

harder for them to successfully establish an infection [72]. The occurrence of antibiotic associated *C. difficile* infection as described previously is an example of the protective role of the gut microbiota. Some species produce or initiate the secretion of molecules that exert antimicrobial effects against invading fungi, viruses and other bacteria [40]. For example, research has shown that another commonly isolated gut species, *Bacteroides thetaiotaomicron* induces the secretion of angiogenin; a bactericidial protein specific to the gut [73]. Factors such as this demonstrate the integral importance of the gut microbiota and that we have a circumstantial dependence on their presence. This dependence can be further illustrated in studies of germ free animals whereby the absence of gut colonisers results in an overall decrease in health and an increased susceptibility to infection [74]. Specifically, germ free or gnotobiotic animals have been shown to have reduced digestive enzyme activity, a decrease in mucosal cell turnover, decreased cytokine production due to the decreased exposure to indigenous microbiota and altered structure and morphology of the epithelial wall such as much thinner intestinal wall with thinner and more pointed villi and a reduction in mucosal surface area and smaller peyer's patches resulting in the very little secretoy IgA [75]. After exposure to typical enteric gut bacteria, germ-free animals begin to resemble normal mucosal structure and function [76]. Germ free animal models have also been shown to have an increased intake in calories, eating on average 30% more than conventional animals in order to sustain their body weight [75, 77] suggesting there may also be a role for the indigenous bacterial gut microbiota in fat deposition [78] especially considering gut bacteria are essential for the metabolism of otherwise indigestible dietary components.

1.1.3.3 Allergies

The importance of the protection and development of immunity that exposure and colonisation of commensals gives us has been re-iterated by what is known as 'the hygiene hypothesis' first proposed by work of Strachan [79] which essentially suggested that increased personal cleanliness and reduced exposure to infectious agents was correlated with an increase in hay fever. Subsequent research has linked other allergic responses and even autoimmune diseases to decreased exposure to bacteria [80]. The 'disappearing microbiota' hypothesis put forward by Blaser and colleagues [81] suggests that one of the main factors behind the increase in allergies and metabolic

diseases in the western world is a consequence of a loss in our ancestral microbiota as opposed to the 'hygiene hypothesis'. It is believed that we co-evolved to form our mutually beneficial relationship with our microbiota with global differences such as diet, lifestyle, geography and ecological disturbances as a key driving force leading to out 'indigenous' microbiota. In the 'disappearing microbiota' theory it is believed that our ancestral microbiota is diminishing as a result of a decrease in the vertical transmission of the gut microbiota. Decreases in exposure to certain microbial species via horizontal transmission (water, food etc.) mean bacteria that could have been acquired via either mode of transmission, are not acquired at all [82]. A key example of this phenomena involves the loss of *Helicobacter pylori* colonisation in inhabitants of the developed world [83]. Loss of the species is likely due to decrease in transmission and also increased antibiotic usage and in parallel, peptic ulcer disease and gastric cancer rates are decreasing [84]. However, incidence rates of oesophageal reflux and adenocarcinoma are increasing. The 'disappearing microbiota' hypothesis suggests that the two trends in disease incidence are intricately linked to the colonisation of *H. pylori* [82].

1.1.4 The Gut Microbiota and Disease

Aside from the beneficial functions, the gut microbiota has a dark side and can also be a significant liability. Certain microbial pathogens are also able to invade the human gut thanks to an arsenal of evolved virulence factors.

1.1.4.1 Invasion by pathogens

Infectious intestinal diseases can be caused by a plethora of microorganisms not only belonging to the Bacteria Kingdom and all anatomical parts of the human gut, including the oral cavity, can gain an infection. The main bacterial causative agents of gastrointestinal infection (gastroenteritis) include; *Camplyobacter* spp. *Clostridium difficile,* enterotoxigenic *E. coli, Salmonella*, *Shigella* and *Vibrio cholera.* Often, the consequence of the infection are diarrhoea, vomiting and abdominal pain. Enterotoxigenic bacteria are contracted via external sources, mainly food or contaminated water and are well equipped for adherence to and invasion of intestinal epithelial cells.

Aside from the invasion of specific pathogens, gastrointestinal infectious can also be largely attributed to the consequence of an external stimuli such as dietary changes, illness, use of antibiotics instigating a change in the composition of the gut microbiota [85]. This change is known as microbial dysbiosis and can have serious implications on host health (figure 1.2). Many bacteria residing in the gut are opportunistic pathogens, that upon suppression of the immune system, including suppression of the protective indigenous microbiota, can evade our defences and establish an infective lifecycle leading to gut associated diseases such as diarrhoea and colitis. Another theory implicating dysbiosis in disease is due to a shift in abundance of 'aggressive' commensals normally fewer in abundance but for one reason or another, become enhanced and indirectly cause an alteration in the genetic pool of the gut microbiota with different metabolic functions leading to either a different immune reaction from the host, or different metabolic activities of the gut microbiota which may contribute to the disease. This is one of the theories that is involved in the pathogenesis of Inflammatory Bowel Disease (IBD) and is discussed in much more detail in Chapter 7.0.

Gut microbial dysbiosis has also been linked with a number of other serious illnesses including type 2 diabetes, alcoholic liver disease, necrotising enterocolitis [86] and metabolic syndrome or obesity [47]. Perhaps one of the most well established incidences of an external stimuli altering the gut microbiota and leading to disease is antibiotic-associated diarrhoea notably caused by overgrowth of *Clostridium difficile* as discussed previously.

One of the main questions that has arisen as a result of these associations is whether or not these gut bacteria are directly influencing the pathology of such diseases or whether the changes are just 'collateral damage' as a consequence of the disease, but before this can be determined it is important to first determine the specific phylotypes or genera and possibly even species that are altered in the plethora of disease states associated with dysbiosis of the gut microbiota and also, the genes and functions accompanied that may be implicated in the molecular basis behind the disease.

1.1.4.2 Irritable Bowel Syndrome and Inflammotory Bowel Disorders

Both Irritable Bowel Syndrome (IBS) and Inflammatory bowel disorders (IBD) such as Ulcerative colitis and Crohn's disease are debilitating, chronically recurring disorders of the human gastrointestinal tract. There is clinical overlap between the two diseases [87] and within IBD itself there are numerous forms of disease of which the two main disorders are Crohn's disease (CD) and Ulcerative colitis (UC) though diagnostically different, they share similar disease aetiology which include alterations in the gut immune response [88], inflammation and dysbiosis of the gut microbiota with shared symptoms characterised by abdominal pain, irregular bowel patterns including bouts of diarrhoea and constipation [88]. The cause of such diseases is multifactoral, but can be mainly attributed to; the genotype of the sufferer, environmental stimuli and host-microbe interactions occurring at the gut/epithelial barrier interface. Different substrate availability as a result of the sufferers genotype, different conditions in the gut, proliferation of pathogens due to impaired immune system, overactive or uncontrolled response of the immune system to commensal bacteria are microbial related factors involved in the onset of IBD [89].
The relationship between the gut microbiota and IBD remains complex and is discussed in extensive detail in Chapter 7.0.

1.1.4.3 Cancer

It is well established that infection by viruses, bacteria or eukaryotic microorganisms are involved in the development of malignant tumours in humans, for example the epstein barr virus can cause lymphomas and nasopharyngeal cancer [90] the human papillomavirus causes cervical cancer [91], and relevant to the human gut; *Helicobacter pylori* is known to cause gastric cancer [92]. Similar to the case of IBD, the relationship between bacteria and colon cancer is complex and as yet, far from established. Colon cancer is the fourth most common cancer in the UK and is the second leading cause of death due to cancer in developing countries [93]. Like IBD, it is also considered a disease of the developed world [94].

Colon cancer's onset is due to a series of somatic mutations [94] and develops after the initiation of colonic polyps. However not all polyps develop into malignant tumours

suggesting that other stimuli are involved in the development of cancer, one of which is potentially the gut microbiota.

However current lab techniques are inadequate to support the culture of most of the gut microbiota, and so studying their interaction with the host, and their involvement in diseases such as colon cancer, is no trivial challenge.

1.1.4.4 Obesity and related metabolic disorders

Obesity has become prolific in the westernised world substantially caused by excessive caloric intake and is characterised by excess body fat. It is also associated with numerous other diseases including heart disease, metabolic endotoxemia and diabetes. At present, the disease poses a significant threat to many inhabitants of the developed world. While disproportionate caloric intake is obviously the main cause of obesity, there is evidence to suggest a genetic predisposition to the disease including altered microbial diversity and subsequent metabolic function.

Research by Ley *et al.*, have demonstrated altered gut microbial ecology in obesity in an elegant study using genetically obese mice (mutations in the leptin gene [95]) and lean littermate comparisons and found significant taxonomic difference between the two dominant phyla. *Firmicutes* were markedly higher in the obese group while the *Bacteroidetes* phylum was diminished. The implications for metabolic function of the gut are enrichment in enzymes involved in degradation of indigestible polysaccharides and have an increased ability to harvest energy from dietary components [9]. This research coupled with other research that has shown enhanced adiposity as a result of the transplant of a gut microbiota acquired following a diet high in fat and low in fibre [47] in mouse models provide evidence for a causative role of the gut microbiota in the disease. Backhed *et al.*,[96] showed that following inoculation of germ-free mice with the gut microbiota of conventional mice, the recently inoculated mice incurred a 60% increase in fat accumulation after just 2 weeks despite a reduced diet and an enhanced exercise routine [96]. The gut microbiota in obesity in humans has also been studied particularly with the use of twin studies and concordant variations in phyla have been observed [41]. Type 2 diabetes (T2D) is characterised by hypoergylcemia and insulin resistance and appears to be intricately linked with obesity and people that are obese are

at considerable risk of developing T2D. Dysbiosis of the gut microbiota has been observed in patients with diabetes [97]. Obesity and diabetes can be characterised by low-grade gut inflammation. The altered gut microbiota is thought to affect homeostasis of glucose and lipids contributing to insulin resistance and hyperglycemia. Metabolic endotoxemia is also a contributory factor toward obesity and diabetes progression. Metabolic endotoxemia occurs as a result of increased levels of lipopolysaccharide (LPS) resulting in a deregulated inflammatory episodes within the gut, enhanced gut epithelial permeability and insulin resistance [98] and consequently, it has been postulated that the gut microbiota in obesity may alter the pool of luminal LPs, causing intestinal inflammation and increased permeability and contributes to insulin resistance and the onset of T2D.

1.1.5 "The Unculturable majority" and culture independent technqiues to study the gut microbiota

Until relatively recently, bacterial culture and microscopic observation were the ultimate means of characterising microorganisms and understanding microbial composition of environmental habitats. For example, the human gut microbiota was first observed by these methods and revealed 113 different bacteria present [99]. While such culture dependent techniques still provide an invaluable means of accessing the physiology and biochemistry of microorganisms, the advent of culture-independent techniques has allowed microbiologists and geneticist to observe the full extent of microbial diversity and abundance. It is now firmly understood that the bacteria isolated via culturing an environmental sample are rarely the most numerically dominant *in situ* [100].Those that do grow so successfully, often inhibiting the growth of others as well, are commonly referred to as 'weeds' This culture bias has become known as 'The great plate count anomaly'[101]. Culture-independent techniques have become the more desirable path to take to analyse microbial community diversity and function. Profiling of microbial communities can be characterised via denaturing gradient gel electrophoresis (DGGE) analysis of the 16S rRNA gene and has been implemented to study bacterial communities of the gut [102]. Other analyses that have been implemented to study gut microbial communities include terminal restriction fragment length polymorphism (T-RFLP) [103], fluorescence *in situ* hybridisation (FISH) and DNA microarrays [104]. These techniques are relatively high-throughput and cost effective, however, it is the use

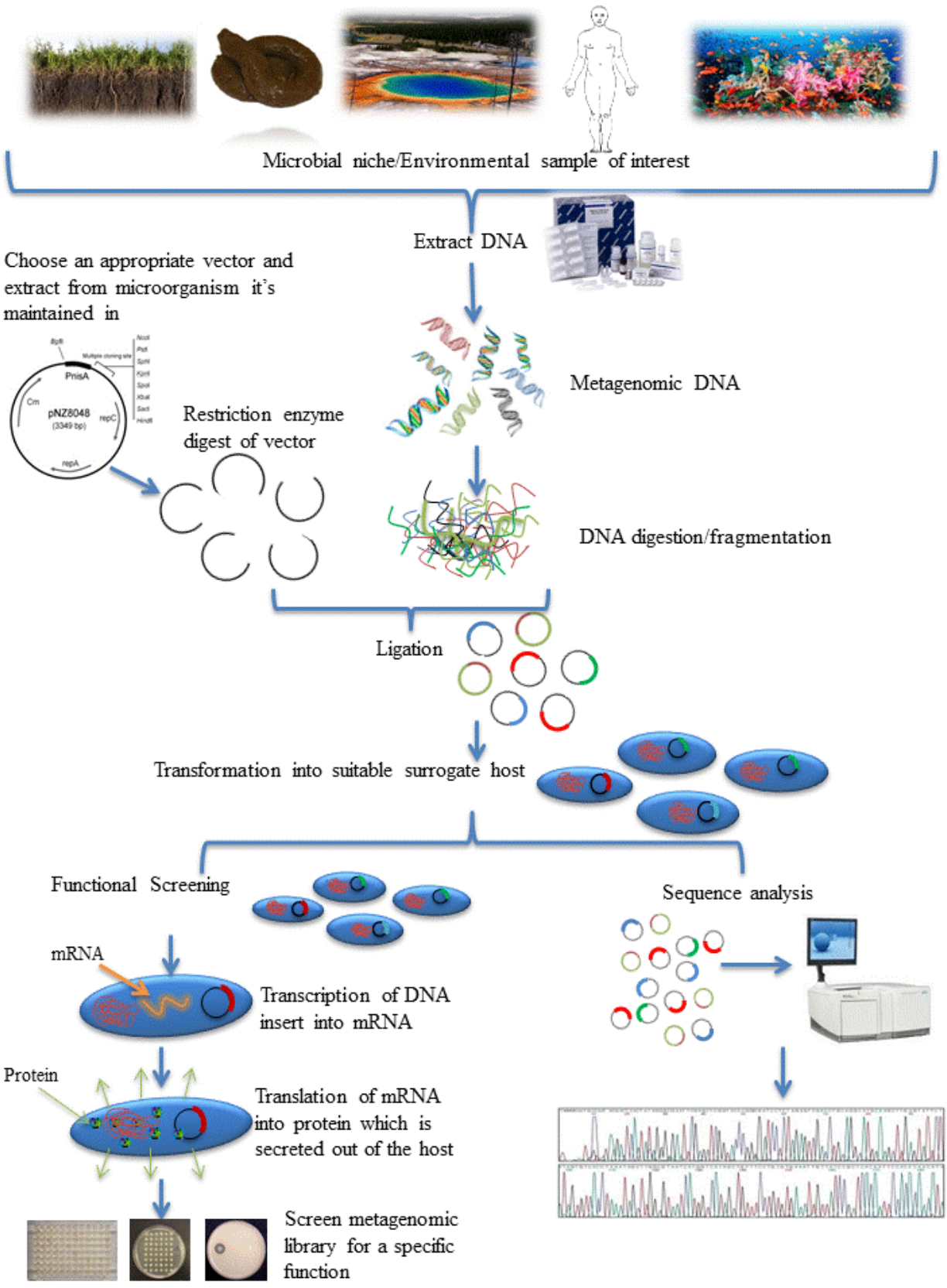
of next generation sequencing of 16S rRNA amplicons through various culture-independent techniques that is becoming the ‘gold-standard’ for studying complex microbial communities (particularly for studying the gut microbiota) due to their extremely high-throughput, decreasing costs and the enormity of data generated from the techniques.

1.2 Metagenomics

Arguably, the emergence of metagenomics in recent years represents the most momentous development in the field of microbial genomics and ecology. Metagenomics has become one of the most valuable ‘go-to’ technologies for studying microbial communities as they are in their natural environment and has significantly aided the progression of culture-independent technology in over-coming “the great plate count anomaly”. Metagenomics can be defined as a direct ‘meta’-analysis of the genomes of all the micro-organisms isolated from a particular environment, bypassing the need to culture [105]. Analysis of genetic information can be conducted in a manner analogous to that of studying single genomes isolated from cultured microorgansisms to elucidate microbial diversity, abundance and key genetic functions within a microbial community. As mentioned previously, it was the implementation of metagenomics and NGS that has allowed us to elucidate both the enormity of diversity and abundance within the gut microbiota [2, 4, 23, 28].

The term ‘metagenomics’ was first coined in 1998 and was implemented to study the soil microbiome by cloning environmental DNA [106], but the basic principles underlying the concept of metagenomics; studying the entire genomes of a mixed community of bacteria taken from an environment, had been used before by Norman Pace’s lab [107] with the understanding that microorganisms of a particular environment are all intricately linked and interact with one another and should therefore be viewed as a functioning interrelated system rather than simply individual organisms.

The basic principles and procedures behind metagenomics are depicted in figure 1.3.


Microbial niche/Environmental sample of interest
Extract DNA
Choose an appropriate vector and extract from microorganism it's maintained in
PnisA
pNZ8048
(3349 bp)
repC
repA
Cm
Restriction enzyme digest of vector
Metagenomic DNA
DNA digestion/fragmentation
Ligation
Transformation into suitable surrogate host
Functional Screening
mRNA
Transcription of DNA insert into mRNA
Protein
Translation of mRNA into protein which is secreted out of the host
Screen metagenomic library for a specific function
Sequence analysis

← *from previous page*

Figure 1.3 A schematic diagram of the general methodology encompassing functional metagenomics and sequencing metagenomics. Upon identification and validation of a positive clone via FM, the insert will be sequenced and subject to extensive bioinformatic analysis also. The sequencing analysis pipeline is not included but is explained in more detail in the text.

The mains steps in the process consist of isolating DNA from a given environment and cloning random fragments into an appropriate vector which is subsequently transformed into a surrogate host for construction of the metagenomic library. [108].The library represents portions of the genetic repertoire of the microbial community within the environment in question and can be analysed by one, or a combination of two methods: sequencing analysis or functional analysis. These techniques will be discussed in more detail but in the first instance, both techniques rely on the isolation of environmental DNA.

1.2.1 Sample Preparation and DNA Isolation

The extraction of total DNA is relevant to both sequencing and functional metagenomics and following experimental design, one of the most crucial steps in metagenomics studies is the isolation of good quality DNA. As highlighted in a review by Ekkers et al., the DNA extraction protocols need to results in genomic DNA of high yield, high DNA quality/purity, large fragment size, and give an unbiased representation of the microbial community in question which can be difficult as some bacterial taxa, such as Gram positive organisms, require harsher lysis methods to recover high yields of DNA so it may be more difficult to isolate less abundant bacteria. Often, enhancement of one such criteria results in negative effect on one of the other factors [109]. In order to achieve a fair representation of the microbial community under scrutiny, DNA can be normalised based on G-C content. This normalisation can be achieved by caesium chloride gradient centrifugation in the presence of an intercalating agent (normally a dye) which will preferentially bind to specific regions of the DNA (for example biz-benzimide is often the intercalating agent and this dye preferentially binds to AT regions). In the presence of the caesium chloride gradient, the DNA will fractionate allowing separation of the

DNA sample based on base content Short *et al*., (1999) DNA can also be normalised by denaturing DNA fragments. Here, the environmental DNA sample is denatured and left to re-anneal to double stranded DNA once more under strict conditions for an extended period of time (commonly 68 °C for up to 36 hours). DNA from abundant species will anneal faster than the less abundant and upon re-annealing, the DNA is separated allowing the less abundant DNA to re-anneal and become enriched [110].

Genomic DNA can be extracted from any sample type for use in FM analyses. However, cell lysis and consequently the DNA extraction of the entire microbial community can be carried out either directly from the environmental sample in question, or by indirect means where cells are first isolated from the sample (for example, by Nycodenz density gradient [111]) and subsequently subject to cell lysis and DNA extraction. Cell lysis may be carried out similarly for both direct and indirect methods and can include mechanical (e.g., bead beating), enzymatic, or chemical-based cell breakage. Although bead beating can yield heavily sheared DNA which can be challenging to generate 'sticky ends' for subsequent ligation, blunt end cloning is usually more successful when working with heavily sheared DNA [110]. A number of studies have been conducted to determine optimum DNA extraction methods for different environments. Salonen and co-workers compared 4 methods for isolating DNA from fecal samples, for metagenomic studies of the human gut microbiota. The four different methods entailed; differential centrifugation and enzymatic lysis, the Promega Wizard DNA extraction kit relying purely on enzymatic lysis, repeated bead beating rounds with SDS, salt and EDTA and lastly, the QiaAmp DNA stool mini kit coupled with a preceding bead beating step [112]. They found the extraction methods to give highly variable DNA yields with as much as 35-fold variations in DNA yields amongst the methods. However their results demonstrated that optimal DNA recovery from human faecal samples was as a consequence of coupling mechanical DNA extraction by repeated bead beating followed by precipitation [112]. Comprehensive and copious studies have been conducted on comparison of DNA extraction methods on different types of soil sample since this environment is often a site dedicated to FM screening for antibiotics [113] and due to the sheer enormity of the soil metagenome [114]. Purohit and Singh assessed 'soft' lysis (SDS and enzymatic extraction), 'harsh' lysis (bead beating and sonicator) and soft and harsh together DNA extraction methods for isolation of DNA

from saline soil habitats. They found that coupling of bead beating with the soft lysis method provided the purest quality, high molecular weight DNA [23-26].

Once the extracted DNA has been verified to be of sufficient quantity, the more the better but most methodologies specify a DNA concentration range that should be expected after correct extraction procedure. The quality of the DNA can be tested to verify that no inhibitors have been co-extracted, DNA is typically tested using an enzymatic test such as PCR amplification with broad-spectrum primers (e.g. primers targeting conserved regions of bacterial ribosomal RNA genes). Extracted DNA is also run on a gel to ensure high molecular weight DNA has been extracted. Once verified, DNA can be manipulated for further steps of functional metagenomic analysis namely ligation into an appropriate vector (size of DNA is also dependent on the type of vector being used) and transformed into a surrogate host for subsequent screening.

1.2.2 Sequencing-based metagenomics

The process of metagenomic sequencing, coupled with the development of next generation sequencing technologies has paved the way for perhaps the most significant developments in microbial genomics. It has allowed the deduction of entire microbial genomes, microbial diversity, abundance and adaptation to environments and the mapping of entire microbial communities [115]. Sequencing metagenomics traditionally involves large DNA inserts, size selected by pulse-field gel electrophoresis and a vector that is capable of maintaining a large DNA insert such as a bacterial artificial chromosome (BAC) cosmids or, most commonly, a fosmid which must be packaged into a phage and transfected into a host. This metagenomic library can be sequenced to identify microbial diversity, abundance and gene content as well as achieving an overall characterisation of the microbial environment in question. Sequencing strategies originated from Sanger sequencing which is still considered a ‘gold standard’ process for sequencing due to its low error rate and relatively long read length of over 700 bp [116], but huge developments in next generation sequencing technologies (NGS) such as 454 sequencing or Illumina (discussed further in 1.3.3) have enabled the generation of extremely large datasets consisting of millions of gene sequences inevitably containing novel genetic information. Sequencing technology is a particularly dynamic area of research at present. Techniques are constantly progressing at costs that are constantly

decreasing making next generation sequencing a possibility for many laboratories globally. NGS methodology includes the preparation of genomic or metagenomic DNA under scrutiny, a library is prepared small synthetic oligonucleotide adapters are ligated to the ends of the fragments. Fragments are amplified by PCR and prepared in a particular way specific to the NGS platform being used. DNA libraries are sequenced and imaged and subject to extensive data analysis [117]. Some of the most commonly used NGS platforms include the Life Technologies Ion Torrent Personal Genome Machine (PGM), the Illumiona MiSeq, Roche's 454 pyrosequencing, Applied biosystems SOLiD system and the PacBio RS II system. Sequencing throughput has increased at an unprecedented rate in recent years, and we are currently able to acquire gigabases of sequence data in a few days [118]. Following retrieval of such data, it must be processed to remove poor quality sequence, raw data as a result of the sequencing technology used (such as barcode or ligated adaptor sequence), vector screening using programs such as cross_match (www.phrao.org) VecScreen (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html) and LUCY [119]. Contaminant removal is also necessary but no trivial task, namely for *E. coli* contamination (as *E. coli* is routinely the cloning host) [105] or screening for human (host) contamination. Sequences are assembled into contigs (contiguous stretches of DNA) via *de novo* sequence assembly or alignment using the increasing variety of bioinformatics software [120] [118]. Bioinformatic gene prediction and annotation follow a lengthy assembly process. Genes can be predicted by homology based searches using the BLAST algorithm [121] or by an '*ab initio*' process to identify genes that do not necessarily have homologs via identification (prediction) of coding regions of DNA. There are a number of programs that assist with this kind of gene prediction such as GeneMark [122], MetaGene [123] Orpheus [124] and GLIMMER [125]. After gene prediction, genes are functionally annotated and predicated by similarity searches to already annotated genes and again, this is usually preformed using function prediction software such as the PFAM database [126], RPS-BLAST [127] and the assignment of clusters of orthologous groups (COGs) [128].

From here, entire genomes or structures of genome communities can be put together and microbial diversity deduced [129] as well as functions of communities [115].
One of the first milestone, large scale metagenomics projects implemented the original DNA sequencing technology of Sanger sequencing to characterise the Sargasso sea

microbiota. In this project, > 1 billion bp of DNA were sequenced and over 1.2 million novel genes were identified including over 782 rhodopsin-like photoreceptor genes cementing a recent suggestion of a level of phototrophic function within marine environments [130], a function previously believed to be exclusive to halophilic archaea [131] .

Metagenomics is becoming a baseline technology for understanding the diversity of microbial ecosystems, and from here, further hypotheses can be developed and with an ever decreasing cost of sequencing technologies due to such a significant amount of competition, sequencing metagenomics is becoming possible for many laboratories worldwide [100] and so is becoming a nascent technology for many microbial genetics research. However, there are also some important limitations associated with this sequence-based approach that must not be overlooked. Firstly, sequencing metagenomics relies entirely on identification of sequence similarity with already known sequences which have been uploaded to public databases, and sequences that are not already known are left unidentified meaning that the entire 'picture' cannot be interpreted and entirely novel genes will rarely be identified. Also, many of the reference databases are currently extremely bias, dominated by just 3 phyla; *Proteobacteria, Firmicutes* and *Actinobacteria* [105]. Secondly, entire gene sequences are rarely revealed using this approach meaning there is always an aspect of ambiguity surrounding deduced gene products preventing a comprehensive biochemical and functional analysis; you can never be positive about the actual function of the gene sequence, additionally, small genes are often missed out completely. At present, it is probably fair to say that the resources and tools available to aid in the functional annotation process are insufficient to keep up with the demands of the huge amount of metagenomic data being produced [132]. Therefore some form of phenotypic analysis is the only alternative to find novel genes and to be able to understand and characterise their function.

1.2.3 Functional metagenomics

Functional metagenomics (FM) is a strategy to identify novel functional genes. FM relies on screening clones in these metagenomic libraries for a particular desired function that is exemplified by heterologous gene expression of the randomly inserted

DNA. Expression of a function can be indicated by various means, namely a measurable phenotype. These screening techniques can include observable phenotypic morphological or physiological changes such as alterations in colony pigmentation, size, shape or growth (inhibition). The use of indicator dyes is also indicative of the expression of particular genes via enzymatic reaction. To date, the majority of successful FM projects identify hydrolytic enzymes such as lipases, esterase and glycoside hydrolases (see table 1.2) and this is likely due to the relative ease of screening methodology. The incorporation of BCIP (5-Bromo-4-chloro-3-indolyl phosphate) in the growth media is a commonly used methods that allows for the colorimetric detection of alkaline phosphatase activity as alkaline phosphatases will hydrolyse BCIP to 5-Bromo-4-chloro-3-indole which is subsequently oxidised to form a dark blue dye. Dark blue colonies are thus indicative of a heterologously expressed alkaline phosphatase which may very well be novel.

Other screening means involve the insertion of, and heterologous expression of a novel gene that allows growth of the host organism, and a screening means known as heterologous complementation. For example Culligan *et al.*, isolated novel salt tolerance inferring genes after selection for *E. coli* clones that were able to grow at elevated salt concentration, concentrations at which the host could not grow [133]. Similarly, Simon and colleagues identified novel DNA polymerase genes by utilising a strain of *E. coli* harbouring a mutation in the 5'-3' exonuclease domain of DNA polymerase. Following growth of the metagenomic library at 18 °C clones harbouring an insert encoding functional DNA polymerase would grow only [134].

The progression of high-throughput screening technologies has markedly improved the efficiency with which a metagenomic library can be screened for a particular function. Jones *et al.*, for example, aimed for a library of over 250,000 clones [70] and Culligan *et al.*, screened over 20,000 [133] and this is not abnormal as numbers such as this are needed to generate appropriate levels of genetic coverage for the environment in question . Automating the process of screening by way of microplate readers and in particular, colony picking robot such as the Pickolo™ and the RapidPick™ have enormously aided the ability to screen such large numbers of clones in a timely, repeatable and logical manner. Sophisticated, high-throughput screening methodologies are becoming developed too. Such methods have incorporated the use of high

performance liquid chromatography (HPLC) and the production of chromogenic or fluorescent or bioluminescent substrate. For example Rabausch and colleagues developed the META system (metagenome extract thin-layer chromatography analysis system) for detection of glycosyltransferases [135]. SIGEX (substrate-induced gene expression) and PIGEX (product-induced gene expression) systems have also been developed [136, 137]. Here (SIGEX), a green fluorescent protein (GFP) vector is used to construct the metagenomic library. The library is subject to a gene of interest induction and if the gene of interest if present and can be expressed by the FM host, the downstream GFP will also be expressed and thus the inserted gene can be identified via fluorescence-activated cell sorting (FACS). PIGEX is similar except the GFP reporter gene is located on a product of enzymatic (gene of interest) activity [137]. Other novel screening methodologies recently developed include the METREX system [138].

Table 1.1 Recent successful functional metagenomic projects, the hosts that were used for heterologous expression, the vector, the assay type and the number of clones screened compared to the number of positive hits.

Environment	Target gene	Host(s)	Vector / Average Insert size	# positives / #screened clones	Assay type	Ref
Human gut	Dietary fiber catabolic enzymes	*E. coli*	Fosmid (30-40kbb)	310/704,000	Agar plate assay supplemented with polysaccharides	[139]
Soil (deciduous forest, creek bed and cold desert)	Antimicrobial	*Agrobacterium tumefaciens Burkholderia graminis Caulobacter vibrioides* *E. coli* *Pseudomonas putida Ralstonia metallidurans*.	Cosmid (n/a[a])	170,000 (forest) 450,000(creek) 130,000(desert)	Primary agar plate assay screen looking for alterations in colony morphology and pigmentation, overlay with *B. Subtilis* to screen for growth inhibition	[140]
Activated sludge from a coke Waste treatment plant	Amidase	*E. coli* harbouring pCmGFPbenR (benzoate-responsive sensor plasmid	Fosmid (33 kb)	4/96,000	PIGEX	[137]
Soil	4'-phosphoantetheinyl	*E. coli*	Plasmid (2-6 kb)	4/3 x 10^6	Agar plate assay: Blue coloration	[141]
Glacier ice	DNA polymerase	*E. coli*	Plasmid (4kb) fosmid (36kb)	Plasmid = 230,00 Fosmid=4000	Agar plate assay-ability to grow at 18°C	[134]
Compost soil	Cellulase	*E. coli*	Cosmid (33 kb)	4/100,000	Plate assay	[142]
Forest soil and grassland soil	Lipase	*E. coli*	Plasmid& Fosmid	28/2217648 (plasmid) 9/711794 (fosmid)	Agar plate assay-agar supplemented with 1% tributyrin	[143]

Soil from an urban environment	Antibiotic resistance	*E. coli*	Plasmid (2 kb)	39/n/a	Agar plate assay- using inhibitory concentrations of antibiotics	[144]
Soil; agricultural field soil and soil in vicinity of *Mammillaria carnea*	Antibiotic resistance	*E. coli*	Plasmid (6.5 - 7 kb)	11/550,000	Agar plate assay- using inhibitory concentrations of antibiotics	[113]
Human gut	Salt tolerance	*E. coli*	Fosmid (~40 kb)	47[b]/23,040	Agar plate assay-inhibitory concentrations of NaCl	[133] [145][c]
Marine sponge	Antimicrobial	*E. coli*	Fosmid (~40 kb)	1/250,000	Agar plate assay; overlay with *B. cereus*	[146]
Alaskan soil	Antibiotic resistance	*E. coli*	Plasmid Fosmid (~40 kb)	1/(13,201 Mb DNA)	Agar plate assay; inhibitory concentrations of antibiotics	[147]
Deep-sea sediment	Lipase	*E. coli*	Fosmid (15-33 kb)	/681,100	Agar plate assay;agar supplemented with 1% tributyrin	[148]
Unvegetated Antarctic soil	Cellulase	*E. coli*	BAC (5.1 kb)	11/124,000	Agar plate assay; agar supplemented with cellulose	[149]
Human Gut (fecal and ileum mucosa)	Hydrolytic enzymes-(probiotic breakdown)	*E. coli*	Fosmid	Fecal=11 hits Ileum = 49	Agar plate assay; supplemented with carbon source	[150]
Activated sludge from a paper mill	Esterase	*E. coli*	Plasmid (3 kb)	2/40,000	Agar plate assay-agar supplemented with 1% tributyrin	[151]

Forest soil	Protease	*E. coli*	Plasmid (7-12 kb)	1 positive clone	Agar plate assay: agar supplemented with AZCL-casein	[152]
Human gut	Antibiotic resistance	*E. coli*	Fosmid (30 kb)	17/415,000	Agar plate assay- using inhibitory concentrations of antibiotics	[153]
Biomass from sequencing fed-batch reactor	Lipase Esterase protease	*E. coli* *Strepytomycse lividans*	Cosmid	17/2000 (esterase/lipase	Agar plate assay-agar supplemented with 1% Tributyrin or skimmed milk	[154]
Desert sand (Death valley and Gobi)	Protease	*E. coli*	Plasmid (6 kb) Fosmid (32 kb)	Plasmid 1/30,000 Fosmid 4/17,000	Agar plate assay-agar supplemented with skimmed milk	[155]
Apple Orchid soil	Antibiotic resistance	*E. coli*	Fosmid (30 kb)	13/446,000	Agar plate assay- using inhibitory concentrations of antibiotics	[156]
Elephant faeces and river sediment	Flavonoid-modifying enzymes	*E. coli*	Fosmid (35 kb)	1/50,000	Extra this layer chromatography	[135]
Cotton field	Tannase	*E. coli*	Plasmid (3.5 kb)	1/92,000	Agar plate assay; agar supplemented with X-caprylate	[157]

FM is potentially an extremely useful tool and is the only way that novel genes can be found [108]. It has been shown to have the capabilities of identifying many novel genes, including those of medical significance such as polyketide antibiotic genes [158] and bacterial antibiotic resistance genes [159]. Functional metagenomics has revealed the extent and diversity of CAZymes [160], bile salt hydrolases [70] and a number of studies have revealed numerous degradative enzymes including esterases, glycoside hydrolases and lipases (see Table 1.1). These findings may have useful implications for biotechnology; they may have clinical significance and above all give us a specific insight into the roles these microbes may be playing in the environments they inhabit, an insight that without functional metagenomics, would remain unseen. This process therefore shows potential for not only furthering our knowledge of the microbial world, but also for effective drug discovery, and unravelling resistance mechanisms and potential mechanisms of disease.

However, although FM is a very powerful technique, it is far more labour intensive process than that of the sequencing approach and the frequency of positive results, i.e. an active clone, remains relatively low. To get enough genetic coverage, clone libraries of tens of thousands of clones have to be screened for a positive phenotype.
For a successful metagenomic screen a gene has to be successfully ligated into a vector, transformed into a surrogate host where it must be faithfully transcribed and translated into the fully functioning protein or product which must be secreted from the cell and its function demonstrated [108]. The major issues surrounding the success of FM have been highlighted by Ekkers *et al.*, [161]. The efficiency and yield of DNA extraction from a given environment remains a significant consideration, although this has been discussed in-depth previously, and this is also a consideration in sequencing-based metagenomics. One of the most hampering processes in FM is that it requires heterologous gene expression of foreign DNA in a surrogate, domesticated host. Therefore, careful consideration must be given to; the size on the insert DNA, the cloning vector and importantly, the surrogate host. While *E. coli* is normally the organism of choice for routine cloning procedures, this cannot be the case with FM when the whole point is to clone DNA from potentially unusual, 'exotic', uncultured microorganisms to rely entirely on *E. coli* would be counterintuitive [108]. Consequently, more recently FM is veering towards alternative hosts for more efficient screening, the wider genetic capacity a host system has, the more chance it will be able

to dependably express a foreign gene. Problems with gene expression arise for a number of reasons for example, incorrect protein folding, enzymatic breakdown of product, the product may exert a toxic effect on the host, gene promoter regions may not be recognised (*cis*-acting DNA factors), the formation of inclusion bodies or missing initiation factors or the inability to process signal peptides for secretion of the product *(trans*-acting DNA factors) [161]. Clearly there are a profusion of factors to consider, broadening host range is an attractive response. However, this does limit the vector choice since the chosen vector has to be replicable inside the chosen host and as most fosmids, cosmids and Bacterial Artificial Chromosomes (BACs) are designed for *E. coli* replication, in most instances with alternative hosts FM is restricted to smaller (<15kb) insert libraries using plasmid vectors. Libraries requiring larger insert sizes (exogenous DNA >15 kb) are typically processed using cosmids (15-35 kb), fosmids (25-45 kb) or BACs (100-200 kb). Large insert size libraries have their advantage in that greater coverage can be achieved with fewer clones and increase the likelihood of contained an intact gene or multiple genes involved in specific biochemical pathways.

The implementation of alternative hosts or hosts used in conjunction with *E. coli* for generation of FM libraries can be seen in table 1.1. Multiple host cloning systems implementing a minimum of two hosts (often *E. coli* remains the surrogate of choice for library maintenance however) meaning a diversification of genetic arsenal and thus an increased chance of expression and secretion of functional product. If a 'positive hit' is observed and it is certain that the clone is not a contaminant, the clone is isolated, and the potential novel insert is sequenced. After the sequence of the insert has been determined it is subject to bioinformatic analysis and activity based assays to characterise the expressed molecule. Following sequencing, the sequence is annotated using sequence editing software such as the commercially available Lasergene99 core suite (DNASTAR, Inc., Madison, Wisconsin, USA) or sequence editor modules within the CodonCode Alignment Software (CodonCode Corporation, Dedham, Massachusetts) or free software which includes BioEdit [162] or Mega (Molecular Evolutionary Genetics Analysis) [163] and predictions of open reading frames, structures, promoters and homology searches can be conducted using software such as the NCBI Open Reading frame Finder. For specific enzymes, there are numerous BLAST algorithm based databases publicly available that specialise in specific classes of enzymes for homology searches and structural and biochemical properties

predictions. For example, the MEROPS database for proteolytic enzymes (http://merops.sanger.ac.uk/index.shtml) can be used for homology searches and to determine substrates and inhibitors of proteases [164], the Carbohydrate-Active Enzyme Database can be used to analyse Carbohyrdate-Active Enzymes (CAZymes) such as glycoside hydrolases (http://www.cazy.org/) [165] and the Lipase Engineering Database (http://www.led.uni-stuttgart.de/) [166] can be used for similar analyses on putative lipase enzymes. A comprehensive list of enzyme databases are discussed in the work of Schomburg and Schomburg [167].

As important as the bioinformatic analyses is, the activity-based characterisation of the insert DNA is as important. Often, following bioinformatic analysis, primers can be designed to span the specific coding region of DNA of which the product can be sub-cloned using standard sub-cloning vector systems (such a pBluescript variations or pET vector expression systems) and screened for recapitulation of the primary phenotype. It clones to recapitulate the particular phenotype it can be fair to assume that a gene encoding a functioning protein has been captured via the FM techniques. From here, further functions of the protein can be explored to determine biochemical traits such as temperature optima, pH optima, the effect of certain inhibitors and substrate specificity.

Thus, this amplifies the notion that FM is currently our only strategy for extracting novel biocatalysts and biomolecules and enables their direct characterisation and insights into microbial function while with sequencing approach these parameters can only be conjectured.

1.3 Microbial Proteases: The Degradome

Proteases are responsible for the hydrolysis of the peptide bond between amino acids in proteins and are essential in many biological processes [168] in all organisms. They comprise an extremely diverse group of enzymes and are classified into the main types according to their catalytic type, or the amino acid that is present at their active site giving the groups; aspartic, cysteine-, threonine-, serine-, glutamic- and Metallo-proteases. Proteases act upon proteases via hydrolysis reaction of the covalent peptide

bond in protein and the different proteases are specific in their mechanism of action and the peptide bond that they hydrolyse. One of the first proteases to be studied in detail was Chymotrypsin, a serine protease secreted by the pancreas in humans and gave great insight into the mechanism of action of such enzymes. The clarification of the 3 dimensional structure of the enzyme helped reveal that the arrangement of amino acids in the 3D structure was such that protons and charge could move within the active site of the enzyme allowing amino acids to act as proton acceptors and donors, nucleophiles and electrophiles. For the different type of proteases, different amino acid residues act as the nucleophile to attack the carbonyl group of the peptide bond. For example in Chymotrypsin, a serine residue in the active site partakes in the nucleophillic attack [169], in cysteine proteases, a nucleophilic cysteine thiol is the nucleophile that initiates the proteolytic reaction, for metalloproteases, a metal ion, commonly a zinc ion, is needed for catalysis.

All groups contain proteases that have a broad range of substrates, for example the serine protease group contains Proteinase K, produced by the fungus *Tritirachium album*, which acts on vast array of substrates including keratin, casein and haemoglobin [170]. In contrast, there are many proteases that have a highly specific substrate usually involved in processes that need to be tightly regulated for example small ubiquitin-like modifier (SUMO) proteases [171, 172]. While proteases can be structurally similar, and be members of the same protease type, their difference in substrate specificity is due to their recognition of different amino acid residues within a substrate.

Proteases from all organisms are also implicated in a multiplicity of diseases including Alzheimer's [173], cancer [174], the pathogenesis of HIV [175] and a plethora of bacterial diseases which will be discussed in more detail later. Though, as previously mentioned, proteases are ubiquitous and essential in all living organisms, those produced by the microbial world are often our principle interest due to their relative ease of genetic manipulation, because of certain physiological and biochemical traits, their vast genetic diversity and their roles in virulence [176].

1.3.1 Microbial Proteases and their uses

Due to their genetic diversity, broad substrate specificity, the relative ease it takes to genetically manipulate them to tailor specific needs, microbial proteases are some of the most important and cost effective biocatalysts involved in the production process of a huge number of commercially and pharmaceutically important products [176, 177]. Perhaps some of the most well-known products of which microbial proteases are essential in the production include; laundry detergents, where most notably subtilisins, a group of serine proteases isolated from the *Bacillus* Genus are used to degrade substrate found in certain stains [178], bacterial and fungal keratinases and elastases are used extensively in the leather making industry for bating and unhairing of product [179]. They are extensively used in the food and drink industry; yeast proteases are used extensively in bread baking and beer manufacturing, bacterial and fungal proteases are important in the development of coffee and cocoa and microbial rennets are being increasingly used in the dairy industry [176]. In the pharmaceutical industry, Clostridial Collagenase, a metalloproteases isolated from *Clostridium histolyticum* has many clinical applications. It is used in conjunction with antibiotics to promote wound healing [180] and in the treatment of peyronie's disease in men [181]. Aspraginase, a protease isolated from two bacteria; *E. coli* and *Erwinia chrysanthemi*, is commonly used to treat malignant lymphomas by converting circulating asparginine (for which leukaemia cells are auxotrophic, but normal cells are not) into aspartic acid [182].

1.3.2 Proteases as virulence factors

In addition to their application in industry, microbial proteases are also clinically relevant as proteases are thought to account for up to 5% of the genome of infectious organisms and are often associated with the pathology of certain human diseases acting as significant virulence factors aiding the onset of infection [183, 184]. Recent research has suggested this number may be much higher [168] and with the advance of culture-independent techniques for studying microbial communities it is fair to theorize that there are many proteases, as well many other molecules, that are yet to be discovered many in microorganisms that we are currently unable to culture.

Bacterial proteases may be implicated as virulence factors for a variety of reasons. A virulence factor can be defined as it enables a microorganism to invade, replicate and establish themselves in a host and cause disease by subverting the host immune system. There are many ways that pathogenic bacteria utilise proteases to enable such processes. For example, some bacterial proteases directly act as toxins targeting proteins of whose degradation directly leads to disease, the role of many proteases of pathogenic organisms is to degrade specific proteins of the immune response thus actively inhibiting the immune system, and some proteases may indirectly act as virulence factors by utilising a substrate for nutritional purpose but causing damage to the host as a consequence. Table 1.2 shows some pathogenic bacteria, proteases they produce, and a more detailed description of the mechanism behind how they act as virulence factors.

Table 1.2. Some well known pathogens that utilise proteases as virulence factors in the onset and protgresion of infection.The mechansims of those proteases in disease are also described with the reference that carried out the research to implicate the proteases as significant viruelnce factors.

Species	Protease	Target/function	Reference
Staphylococcus aureus	Metalloprotease aurolysin	Inhibition of host immune response-Blocks phagocytosis and neutrophil activation by cleaving and inactivating complement protein C3	[185]
Bacillus anthracis	Metalloprotease anthrax lethal factor	Evasion of host imune system , cleaveage and inactivation of mitogen-activated protein kinase kinases leading to disregulation in many cellular processes such as apoptosis	[186]
Clostridium botulinium	Metalloprotease bontoxilysin	This protease is part of the botulinum toxin that binds and cleaves synaptosomal-associated proteins consequently preventing vesicle fusion and release of the excitory neurotransmitter, acetylcholine.	[187]
Helicobacter pylori	Serine protease HtrA	Cleaves the transmembrane glycoprotein e-cadherin leading to disruption of epithelial barrier function. E-cadherin is also considered a tumour suppressor protein due to it's role in cell-cell adhesion. Loss of funtion of e-cadherin allows the proliferation of tumour cells.	[188]
Pseudomonas aeruginosa	Thermolysin-like metallo protease; Elastase	Elastase B from *P. aeruginosa* has a broad spectrum of activity contributing to it's virulence including disprution of epithelial barrier function by degradation of tight junctions and degradation of a nu,ber of innate and adaptive immunity proteins including TNF-α and IFN-γ, IgA and IgG.	[189]
Enterococcus faecalis	Metalloprotease; Gelatinase E (GelE)	GelE has a broad spectrum of activity which contributes to *E. faecalis* virulence including disruption of epithelial barrier function and evasion of the host immune system by degradation of C3 and iC3b	[190]

Proteus mirabilis	Metalloprotease; ZapA/mirabilysin	ZapA degrades IgA1, IgA2 and IgG	[191]
Bacteroides fragilis	Metalloprotease; Fragilysin	Hydrlyses the extracellular domain of E-cadherin to disrupt epithelial barrier function	[192]
Enterotoxigenic/Enterohemorrhagic E. coli	Metalloprotease StcE	StcE is sectreted by the type II secretion system and degrades the high molecular weight glycoprotein; mucin lining the epithelaial cells of the human gut thus allowing intimate adherence of the E.coli to host cells.	[193]
***Clostridium* spp.**	Collagenase A	Degrade collagen	[194]
Vibrio cholerae	Metalloprotease; TagA	Carries out the same function as StcE	[195]
Mycobacterium tuberculosis	Mycosin-1 subtilisin-like serine protease	Mycosin-1 activates the ESX-1 secretion system for transportation of other protein and virulence factors into infected macrophages and is essential for establishment in initial infection	[196, 197]

1.3.3 Protease activity in the human gut

Given that a large aspect of the human diet is constituted of protein, it is probably unsurprising that there is substantial amount of exogenous (dietary) protein available as substrate for the gut microbiota [198, 199]. Macfarlane and colleagues also demonstrated that the microbiota harbours large quantities of proteases that may be contributing to proteolysis of dietary protein [200].

However, we hypothesise that there may also be a substantial amount of substrate available for degradation by the gut microbiota that is endogenous to the host. Examples of such protein include the high molecular weight glycoprotein mucin, keratin [201], proteins of the immune system including immunoglobulins and antimicrobial peptide and protein that make up tight junctions between intestinal epithelial cells and contribute to gut barrier function (cadherins, zonulin, claudin). The ability of bacteria to degrade these proteins may have implications on host health since in order to develop and maintain themselves in the GI tract, bacteria must compete against both the primary defence including physical barriers such as mucin and the innate immune system such as non-specific secretary Immunoglobin A- the most abundant immunoglobin in the gut, the vast array of other antimicrobial peptides and must outcompete other bacteria for nutrients needed for growth. Bacterial degradation of such proteins may have relevance in the pathological process of a number of gut related disorders.

It was mentioned previously that while there are many contributing factors to disorders such as Crohn's disease and Ulcerative colitis, one of the key processes is host-microbial interactions at the gut barrier interface and previous research has shown that there is elevated microbial proteolytic activity in the gut of those suffering with disorders such as IBS. Increasing evidence for the roles of proteases in disorders of the human gut stem from pinnacle studies such as those conducted by Steck *et al.*, whereby a matrix metalloprotease, *gelE*, isolated from *Entercoccus faecalis*, often a commensal of the gut microbiota, was able to directly contribute to chronic intestinal inflammation to animal models of IBD via impairment epithelial barrier function and increased extracellular permeability through degradation of the transmembrane glycoprotein, E-

cadherin [202]. Prunteanu and colleagues also found that a proteolytic isolate form the gut, *Clostridium perfringens*, was capable of reducing transepithelial resistance in Ussing chamber experiments thus suggesting the ability of this isolate to compromise epithelial barrier function. Two pinnacle reviews have been published outlining the possible mechanisms by which bacterial proteases may be contributing the inflammatory disorders of the gut [203, 204]. Recent research as part of the European Union research project; Intestinal Proteases: Opportunity for Drug Discover (IPODD) Consortium (http://www.ipodd.eu/) has also highlighted that the activity of matrix metalloproteases (MMPs) previously believed to be of host origin, may actually be produced by gut bacteria and more so, this may mean that bacteria could be a source of protease inhibitors for both bacterial and host proteases and could be a source of anti-inflammatory drugs.

A more comprehensive view of the underlying molecular mechanisms and interactions that are occurring here is needed if we are to understand the impact the gut microbiota is having on host health. The problems of culture make this difficult however culture-independent processes may facilitate the analysis of microbial proteases that are produced by the vastly uncultivated gut microbiota and may well be implicated in disease.

1.3.4 Proteases and functional metagenomics

Clearly there is much evidence supporting the role of proteases as virulence factors in disorders of the gut. However, when applying experimental data to the human gut, it is inappropriate to simply analyse pure cultures of isolated microorganisms alone as species such as *E. faecalis* and *C. perfringens* man not be numerically dominant in IBD. Also, little is known about their interaction with other members of the gut microbiota. Additionally, as current culture methods do not support the growth of the vast majority of the gut microbiota, there remains a lot of ambiguity surrounding the proteolytic capacity of the uncultured gut microbiota. Therefore alternative strategies need to be implemented if we are to uncover the entire potential and function of the gut microbiota degradome.

As previously discussed, one of the most effective ways to isolate novels genes and therefore functions, in the case of this research, proteolytic functions, of bacteria from environments such as the gut, FM provides the means to access these functions. Functional metagenomics has been used to screen for and successfully isolate a number of different microbial enzymes that have potential for industrial applications [205] and are clinically relevant [206] and provide generous insights into the functional diversity of bacteria and how they have adapted to their environment [70]. However, when it comes to screening for proteases via functional metagenomics there are a currently a number of limitations that have been highlighted previously (section 1.3.3) and the lack of success of FM screening for proteases are well documented [207-209]. Solutions to these limitations need to be explored if we are to access the entirety of the 'Degradome'.

1.3.5 Protease "Degradomics"

In light of increasing knowledge that proteases are intricately involved, if not in control of in many cellular processes as well as pathological processes, increasing efforts have been implemented to determine means for identifying and characterising proteases. The MEROPS database is a large, comprehensive peptidase database currently including sequence and function information for around 3000 proteases and protease inhibitors and provides an excellent resource for the study of proteases that have already been characterised in one way or another and to study protein that are homologous to those already in the database. Data from this site has also helped generate the estimate that proteases account for at least 2% of protein in all organisms [164]. Additionally, another "–omic" technology has also been coined dedicated to studying the function of proteases, their substrate specificity and their inhibitors; Degradomics [210]. Degradomics aims to encompass genomic and proteomic methodologies dedicated to the identification of proteases (family, type etc), elucidating their function, their endogenous inhibitors and their substrate specificity.

Degradomics will help uncover the biological roles of proteases in the environments they exist be it the human body, soil, the ocean and further characterisation will aid treatments for diseases of which proteases are central. Degradomic studies coupled with improvement of and implementation of other metagenomic analyses will significantly advance our knowledge of microbial proteases in the human gut.

1.4 AIMS AND OBJECTIVES

Proteolytic degradation of both exogenous and endogenous protein has been demonstrated to have implications on host health and the gut microbiota has been shown to significantly contribute to proteolysis in the human colon [200]. It has been estimated the proteases make up a significant percentage of protein within the genomes of both pathogens and commensals [164]. There is also data to suggest contribution of bacterial proteases to IBD [202-204].

Thus far research into proteases and their contribution to gut inflammation have been conducted on pure cultures on laboratory isolates. A coupling of genomic and degradomic techniques is required to elucidate the function/role of gut microbial proteases with regards to the environment they inhabit and to determine their potential involvement in disease. Exploration of these techniques and their implementation to investigate the gut microbiota and its degradome in disease surmises the main objective

The main aims and hypotheses of this project are as follows:-

1. **To develop a simple, yet robust agar to eliminate the occurrence of false positive in functional metagenomic library screening for proteases-**We hypothesise that current culture techniques for proteases lead to false positives (clear halo forming on skimmed milk agar) due to the hydrolysis of lactose in the milk and the subsequent pH drop. Using a fat-free, lactose free substitute should ameliorate this problem. (Chapter 3).

2. **To identify and functionally characterise the active protease gene(s) from a fosmid vector isolated using metagenomics from activated milk waste treatment plant.** (Chapter 4).

3. **To conduct a culture-dependent analysis on protease producers within the gut microbiota.** (Chapter 5).

4. **To develop novel metagenomic tools to obtain characterised proteases from (gut) bacteria and to implement these tools to screen the gut microbiota for protease activity.** We hypothesise that the use of a novel shuttle vector system utilising a Gram negative as well as a Gram positive host that has a demonstrated capacity to express and secrete proteases effectively will be much more efficient in capturing protease genes from the gut metagenome. (Chapter 5).

5. **To determine optimal conditions for faecal protease extraction and storage over time.** (Chapter 6).

6. **To conduct a comprehensive comparison of protease activity in an IBD cohort and compare to a group of healthy volunteers.** It is hypothesised the IBD cohort will harbour an increased level of protease activity as protease activity has been associated with a contribution to inflammation in the human gut. Protease activity has also been found to be elevated in diseases of a similar pathology such as IBS.

7. **To conduct 16S rRNA community profiling analysis for each cohort.** Previous studies have found a dysbiosis of the gut microbiota in IBD. We hypothesise that the IBD cohort will show different microbial abundance and diversity when compared to the healthy cohort as a whole. We also hypothesise that compositional alterations in the gut microbiota will be associated with higher levels of protease activity. (Chapter 7).

8. **To further unravel the potential role of the gut degradome in the virulence of inflammation.** As it has been demonstrated that inhabitants of the gut microbiota are capable of secreting proteases that act as virulence factors, we hypothesise that the proteases isolated from the IBD cohort will have an enhanced ability to behave as virulence factors compared to the healthy cohort due to compositions in the gut microbiota. (Chapter 7).

1.5 REFERENCES

1. Costa, M., Brookes, S.J.H., and Hennig, G.W. (2000). Anatomy and physiology of the enteric nervous system. Gut *47*, iv15-iv19.
2. Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S.R., Nelson, K.E., and Relman, D.A. (2005). Diversity of the human intestinal microbial flora. Science. 2005 Jun 10;308(5728):1635-8. Epub 2005 Apr 14.
3. Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J.A., Bonazzi, V., McEwen, J.E., Wetterstrand, K.A., Deal, C., et al. (2009). The NIH Human Microbiome Project. Genome research *19*, 2317-2323.
4. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59-65.
5. Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., and Gordon, J.I. (2007). The Human Microbiome Project. Nature *449*, 804-810.
6. Cusack, S., and O'Toole, P. (2013). Challenges and implications for biomedical research and intervention studies in older populations: insights from the ELDERMET study. Gastroerontology *59*, 114-121.
7. Lozupone, C.A., Stombaugh, J.I., Gordon, J.I., Jansson, J.K., and Knight, R. Diversity, stability and resilience of the human gut microbiota. (2012). Nature *489*, 220-230.
8. Ley, R.E., Peterson, D.A., and Gordon, J.I. (2006). Ecological and Evolutionary Forces Shaping Microbial Diversity in the Human Intestine. Cell *124*, 837-848.
9. Ley, R.E., Bäckhed, F., Turnbaugh, P., Lozupone, C.A., Knight, R.D., and Gordon, J.I. (2005). Obesity alters gut microbial ecology. Proceedings of the National Academy of Sciences of the United States of America *102*, 11070-11075.
10. Marchesi, J.R.(2011). Human distal gut microbiome. Environmental Microbiology *13*, 3088-3102.
11. Mackie, R.I., Sghir, A., and Gaskins, H.R. (1999). Developmental microbial ecology of the neonatal gastrointestinal tract. The American Journal of Clinical Nutrition *69*, 1035s-1045s.
12. Jiménez, E., Marín, M.L., Martín, R., Odriozola, J.M., Olivares, M., Xaus, J., Fernández, L., and Rodríguez, J.M. (2008). Is meconium from healthy newborns actually sterile? Research in Microbiology *159*, 187-193.
13. Jimenez, E., Fernandez, L., Marin, M.L., Martin, R., Odriozola, J.M., Nueno-Palop, C., Narbad, A., Olivares, M., Xaus, J., and Rodriguez, J.M. (2005). Isolation of commensal bacteria from umbilical cord blood of healthy neonates born by cesarean section. Current Microbiol *51*, 270-274.
14. O'Toole, P.W., and Claesson, M.J. (2010). Gut microbiota: Changes throughout the lifespan from infancy to elderly. International Dairy Journal *20*, 281-291.
15. Isolauri, E. Development of healthy gut microbiota early in life. (2012). Journal of Paediatrics and Child Health *48*, 1-6.
16. Dominguez-Bello, M.G., Blaser, M.J., Ley, R.E., and Knight, R. (2011). Development of the Human Gastrointestinal Microbiota and Insights From High-Throughput Sequencing. Gastroenterology *140*, 1713-1719.

17. Adlerberth, I., Lindberg, E., Aberg, N., Hesselmar, B., Saalman, R., Strannegard, I.L., and Wold, A.E. (2006). Reduced enterobacterial and increased staphylococcal colonization of the infantile bowel: an effect of hygienic lifestyle? Pediatric research *59*, 96-101.
18. Penders, J., Thijs, C., Vink, C., Stelma, F.F., Snijders, B., Kummeling, I., van den Brandt, P.A., and Stobberingh, E.E. (2006). Factors Influencing the Composition of the Intestinal Microbiota in Early Infancy. Pediatrics *118*, 511-521.
19. Palmer, C., Bik, E.M., DiGiulio, D.B., Relman, D.A., and Brown, P.O. (2007). Development of the human infant intestinal microbiota. PLoS Biology *5*, 1556-1573.
20. Zoetendal, E.G., Akkermans, A.D.L., and De Vos, W.M. (1998). Temperature Gradient Gel Electrophoresis Analysis of 16S rRNA from Human Fecal Samples Reveals Stable and Host-Specific Communities of Active Bacteria. Applied and Environmental Microbiology *64*, 3854-3859.
21. Zoetendal, E.G., von Wright, A., Vilpponen-Salmela, T., Ben-Amor, K., Akkermans, A.D.L., and de Vos, W.M. (2002). Mucosa-Associated Bacteria in the Human Gastrointestinal Tract Are Uniformly Distributed along the Colon and Differ from the Community Recovered from Feces. Applied and Environmental Microbiology *68*, 3401-3407.
22. Langendijk, P.S., Schut, F., Jansen, G.J., Raangs, G.C., Kamphuis, G.R., Wilkinson, M.H., and Welling, G.W. (1995). Quantitative fluorescence in situ hybridization of Bifidobacterium spp. with genus-specific 16S rRNA-targeted probes and its application in fecal samples. Applied and Environmental Microbiology *61*, 3069-3075.
23. Gill, S.R., Pop, M., DeBoy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., Gordon, J.I., Relman, D.A., Fraser-Liggett, C.M., and Nelson, K.E. (2006). Metagenomic analysis of the human distal gut microbiome. Science *312*, 1355-1359.
24. Hold, G.L., Pryde, S.E., Russell, V.J., Furrie, E., and Flint, H.J. (2002). Assessment of microbial diversity in human colonic samples by 16S rDNA sequence analysis. FEMS Microbiology Ecol *39*, 33-39.
25. Wang, X., Heazlewood, S.P., Krause, D.O., and Florin, T.H. (2003). Molecular characterization of the microbial species that colonize human ileal and colonic mucosa by using 16S rDNA sequence analysis. J Appl Microbiol *95*, 508-520.
26. Hayashi, H., Sakamoto, M., and Benno, Y. (2002). Phylogenetic analysis of the human gut microbiota using 16S rDNA clone libraries and strictly anaerobic culture-based methods. Microbiology and immunology *46*, 535-548.
27. Costello, E.K., Lauber, C.L., Hamady, M., Fierer, N., Gordon, J.I., and Knight, R. (2009). Bacterial Community Variation in Human Body Habitats Across Space and Time. Science *326*, 1694-1697.
28. Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.-M., et al. (2011). Enterotypes of the human gut microbiome. Nature *473*, 174-180.
29. Hehemann, J.-H., Correc, G., Barbeyron, T., Helbert, W., Czjzek, M., and Michel, G. (2010). Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. Nature *464*, 908-912.
30. Lozupone, C.A., Stombaugh, J.I., Gordon, J.I., Jansson, J.K., and Knight, R. Diversity, stability and resilience of the human gut microbiota. Nature. 2012 Sep 13;489(7415):220-30. doi: 10.1038/Nature11550.

31. Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., et al. (2012). Human gut microbiome viewed across age and geography. Nature *486*, 222-227.
32. Mueller, S., Saunier, K., Hanisch, C., Norin, E., Alm, L., Midtvedt, T., Cresci, A., Silvi, S., Orpianesi, C., Verdenelli, M.C., et al. (2006). Differences in Fecal Microbiota in Different European Study Populations in Relation to Age, Gender, and Country: a Cross-Sectional Study. Applied and Environmental Microbiology *72*, 1027-1033.
33. Biagi, E., Nylund, L., Candela, M., Ostan, R., Bucci, L., Pini, E., Nikkïla, J., Monti, D., Satokari, R., Franceschi, C., et al. (2010). Through Ageing, and Beyond: Gut Microbiota and Inflammatory Status in Seniors and Centenarians. PLoS ONE *5*, e10667.
34. Wang, M., Ahrné, S., Jeppsson, B., and Molin, G. (2005). Comparison of bacterial diversity along the human intestinal tract by direct cloning and sequencing of 16S rRNA genes. FEMS Microbiology Ecology *54*, 219-231.
35. Peña, J.A., McNeil, K., Fox, J.G., and Versalovic, J. (2002). Molecular Evidence of *Helicobacter cinaedi* Organisms in Human Gastric Biopsy Specimens. Journal of Clinical Microbiology *40*, 1511-1513.
36. Roos, S., Engstrand, L., and Jonsson, H. (2005). *Lactobacillus gastricus* sp. nov., *Lactobacillus antri* sp. nov., *Lactobacillus kalixensis* sp. nov. and *Lactobacillus ultunensis* sp. nov., isolated from human stomach mucosa. International Journal of Systematic and Evolutionary Microbiology *55*, 77-82.
37. Lawal, O.O., Rotimi, O., and Okeke, I. (2007). *Helicobacter pylori* in gastroduodenal diseases. Journal of the National Medical Association *99*, 31-34.
38. Kashket, E.R. (1987). Bioenergetics of lactic acid bacteria: cytoplasmic pH and osmotolerance. FEMS Microbiology Letters *46*, 233-244.
39. Bik, E.M., Eckburg, P.B., Gill, S.R., Nelson, K.E., Purdom, E.A., Francois, F., Perez-Perez, G., Blaser, M.J., and Relman, D.A. (2006). Molecular analysis of the bacterial microbiota in the human stomach. Proceedings of the National Academy of Sciences of the United States of America *103*, 732-737.
40. Montalto, M., D'Onofrio, F., Gallo, A., Cazzato, A., and Gasbarrini, G. (2009). Intestinal microbiota and its functions. Digestive and Liver Disease Supplements *3*, 30-34.
41. Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P., et al. (2009). A core gut microbiome in obese and lean twins. Nature *457*, 480-484.
42. Reyes, A., Haynes, M., Hanson, N., Angly, F.E., Heath, A.C., Rohwer, F., and Gordon, J.I. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature *466*, 334-338.
43. Benson, A.K., Kelly, S.A., Legge, R., Ma, F., Low, S.J., Kim, J., Zhang, M., Oh, P.L., Nehrenberg, D., Hua, K., et al. (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. Proceedings of the National Academy of Sciences *107*, 18933-18938.
44. Louis, P., Scott, K.P., Duncan, S.H., and Flint, H.J. (2007). Understanding the effects of diet on bacterial metabolism in the large intestine. Journal of Applied Microbiology *102*, 1197-1208.
45. Turnbaugh, P.J., Ridaura, V.K., Faith, J.J., Rey, F.E., Knight, R., and Gordon, J.I. (2009). The Effect of Diet on the Human Gut Microbiome: A Metagenomic

Analysis in Humanized Gnotobiotic Mice. Science Translational Medicine *1*, 6ra14.
46. David, L.A., Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J.E., Wolfe, B.E., Ling, A.V., Devlin, A.S., Varma, Y., Fischbach, M.A., et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. Nature *505*, 559-563.
47. Turnbaugh, P.J., Bäckhed, F., Fulton, L., and Gordon, J.I. (2008). Diet-Induced Obesity Is Linked to Marked but Reversible Alterations in the Mouse Distal Gut Microbiome. Cell Host & Microbe *3*, 213-223.
48. Wu, G.D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S.A., Bewtra, M., Knights, D., Walters, W.A., Knight, R., et al. (2011). Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. Science *334*, 105-108.
49. Pérez-Cobas, A.E., Gosalbes, M.J., Friedrichs, A., Knecht, H., Artacho, A., Eismann, K., Otto, W., Rojo, D., Bargiela, R., von Bergen, M., et al. (2012). Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. Gut.
50. Jernberg, C., Lofmark, S., Edlund, C., and Jansson, J.K. (2007). Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. ISME J *1*, 56-66.
51. Candela, M., Perna, F., Carnevali, P., Vitali, B., Ciati, R., Gionchetti, P., Rizzello, F., Campieri, M., and Brigidi, P. (2008). Interaction of probiotic Lactobacillus and Bifidobacterium strains with human intestinal epithelial cells: Adhesion properties, competition against enteropathogens and modulation of IL-8 production. International Journal of Food Microbiology *125*, 286-292.
52. Candela, M., Maccaferri, S., Turroni, S., Carnevali, P., and Brigidi, P. (2010). Functional intestinal microbiome, new frontiers in prebiotic design. International Journal of Food Microbiology *140*, 93-101.
53. Manichanh, C., Reeder, J., Gibert, P., Varela, E., Llopis, M., Antolin, M., Guigo, R., Knight, R., and Guarner, F. (2010). Reshaping the gut microbiome with bacterial transplantation and antibiotic intake. Genome research *20*, 1411-1419.
54. Khoruts, A., Dicksved, J., Jansson, J.K., and Sadowsky, M.J. (2010). Changes in the composition of the human fecal microbiome after bacteriotherapy for recurrent Clostridium difficile-associated diarrhea. Journal of clinical gastroenterology *44*, 354-360.
55. Coates, M.E., Ford, J.E., and Harrison, G.F. (1968). Intestinal synthesis of vitamins of the B complex in chicks. British Journal of Nutrition *22*, 493-500.
56. Albert, M.J., Mathan, V.I., and Baker, S.J. (1980). Vitamin B12 synthesis by human small intestinal bacteria. Nature *283*, 781-782.
57. Macfarlane, S., and Macfarlane, G.T. (2003). Regulation of short-chain fatty acid production. Proceedings of the Nutrition Society *62*, 67-72.
58. Bernalier-Donadille, A. (2010). Fermentative metabolism by the human gut microbiota. Gastroentérologie Clinique et Biologique *34*, S16-S22.
59. Kranich, J., Maslowski, K.M., and Mackay, C.R. (2011). Commensal flora and the regulation of inflammatory and autoimmune responses. Seminars in Immunology *23*, 139-145.
60. Wong, J.M., de Souza, R., Kendall, C.W., Emam, A., and Jenkins, D.J. (2006). Colonic health: fermentation and short chain fatty acids. J Clin Gastroenterology. 2006 Mar;40(3):235-43.
61. Tremaroli, V., and Backhed, F. (2012). Functional interactions between the gut microbiota and host metabolism. Nature *489*, 242-249.

62. Duncan, S.H., Hold, G.L., Barcenilla, A., Stewart, C.S., and Flint, H.J. (2002). *Roseburia intestinalis* sp. nov., a novel saccharolytic, butyrate-producing bacterium from human faeces. International Journal of Systematic and Evolutionary Microbiology *52*, 1615-1620.
63. Eeckhaut, V., Machiels, K., Perrier, C., Romero, C., Maes, S., Flahou, B., Steppe, M., Haesebrouck, F., Sas, B., Ducatelle, R., et al. (2013). *Butyricicoccus pullicaecorum* in inflammatory bowel disease. Gut *62*, 1745-1752.
64. Frank, D.N., Robertson, C.E., Hamm, C.M., Kpadeh, Z., Zhang, T., Chen, H., Zhu, W., Sartor, R.B., Boedeker, E.C., Harpaz, N., et al. (2011). Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases. Inflammatory Bowel Dis *17*, 179-184.
65. Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermudez-Humaran, L.G., Gratadoux, J.-J., Blugeon, S., Bridonneau, C., Furet, J.-P., Corthier, G., et al. (2008). *Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. Proceedings of the Natioanll Academy of Science U S A *105*, 16731-16736.
66. Lawrence, T. (2009). The nuclear factor NF-kappaB pathway in inflammation. Cold Spring Harbor perspectives in biology *1*, a001651.
67. Segain, J.P., Raingeard de la Bletiere, D., Bourreille, A., Leray, V., Gervois, N., Rosales, C., Ferrier, L., Bonnet, C., Blottiere, H.M., and Galmiche, J.P. (2000). Butyrate inhibits inflammatory responses through NF kappa B inhibition: implications for Crohn's disease. Gut *47*, 397-403.
68. McGarr, S.E., Ridlon, J.M., and Hylemon, P.B. Diet, anaerobic bacterial metabolism, and colon cancer: a review of the literature. J Clinical Gastroenterol. 2005 Feb;39(2):98-109.
69. Ridlon, J.M., Kang, D.-J., and Hylemon, P.B. (2006). Bile salt biotransformations by human intestinal bacteria. Journal of Lipid Research *47*, 241-259.
70. Jones, B.V., Begley, M.i., Hill, C., Gahan, C.G.M., and Marchesi, J.R. (2008). Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. Proceedings of the National Academy of Sciences *105*, 13580-13585.
71. Begley, M., Hill, C., and Gahan, C.G.M. (2006). Bile Salt Hydrolase Activity in Probiotics. Applied and Environmental Microbiology *72*, 1729-1738.
72. Macpherson, A.J., and Harris, N.L. (2004). Interactions between commensal intestinal bacteria and the immune system. Nature Reviewa Immunology *4*, 478-485.
73. Hooper, L.V., Stappenbeck, T.S., Hong, C.V., and Gordon, J.I. (2003). Angiogenins: a new class of microbicidal proteins involved in innate immunity. Nature Immunology *4*, 269-273.
74. Abrams, G.D., and Bishop, J.E. (1966). Effect of the Normal Microbial Flora on the Resistance of the Small Intestine to Infection. Journal of Bacteriology *92*, 1604-1608.
75. Shanahan, F. (2002). The host–microbe interface within the gut. Best Practice & Research Clinical Gastroenterology *16*, 915-931.
76. Umesaki, Y., Okada, Y., Matsumoto, S., Imaoka, A., and Setoyama, H. (1995). Segmented filamentous bacteria are indigenous intestinal bacteria that activate intraepithelial lymphocytes and induce MHC class II molecules and fucosyl

asialo GM1 glycolipids on the small intestinal epithelial cells in the ex-germ-free mouse. Microbiology and immunology *39*, 555-562.
77. Hooper, L.V., and Gordon, J.I. (2001). Commensal Host-Bacterial Relationships in the Gut. Science *292*, 1115-1118.
78. O'Hara, A.M., and Shanahan, F. (2006). The gut flora as a forgotten organ. EMBO reports *7*, 688-693.
79. Strachan, D.P. (1989). Hay fever, hygiene, and household size. BMJ *299*, 1259-1260.
80. Bach, J.-F.o. (2002). The Effect of Infections on Susceptibility to Autoimmune and Allergic Diseases. New England Journal of Medicine *347*, 911-920.
81. Blaser, M.J. (2006). Who are we? Indigenous microbes and the ecology of human diseases. EMBO reports *7*, 956-960.
82. Blaser, M.J., and Falkow, S. (2009). What are the consequences of the disappearing human microbiota? Nature Reviews Microbiology *7*, 887-894.
83. Harvey, r.f., spence, r.w., lane, j.a., nair, p., murray, l.j., harvey, i.m., and donovan, j. (2002). Relationship between the birth cohort pattern of *Helicobacter pylori* infection and the epidemiology of duodenal ulcer. QJM *95*, 519-525.
84. Howson, C.P., Hiyama, T., and Wynder, E.L. (1986). The decline in gastric cancer: epidemiology of an unplanned triumph. Epidemiologic reviews *8*, 1-27.
85. Shen, D., Liu, C., Xu, R., and Zhang, F. (2012). Human gut microbiota: dysbiosis and manipulation. Frontiers in Cellular and Infectious Microbiology *2*, 123-123.
86. Spor, A., Koren, O., and Ley, R. (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. Nature Reviews Microbiology *9*, 279-290.
87. Bercik, P., Verdu, E.F., and Collins, S.M. (2005). Is irritable bowel syndrome a low-grade inflammatory bowel disease? Gastroenterology clinics of North America *34*, 235-245, vi-vii.
88. Steck, N., Mueller, K., Schemann, M., and Haller, D. (2013). Bacterial proteases in IBD and IBS. Gut.
89. Olivares, M., Laparra, J.M., and Sanz, Y. (2013). Host genotype, intestinal microbiota and inflammatory disorders. Br J Nutr *109 Suppl 2*, 76-80.
90. Raab-Traub, N. Epstein-Barr virus and nasopharyngeal carcinoma. (1992). Seminars on Cancer Biology. 1992 Oct;3(5):297-307.
91. Walboomers, J.M.M., Jacobs, M.V., Manos, M.M., Bosch, F.X., Kummer, J.A., Shah, K.V., Snijders, P.J.F., Peto, J., Meijer, C.J.L.M., and Muñoz, N. (1999). Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. The Journal of Pathology *189*, 12-19.
92. Correa, P. (2003). Bacterial Infections as a Cause of Cancer. Journal of the National Cancer Institute *95*, E3.
93. Scanlan, P.D., Shanahan, F., Clune, Y., Collins, J.K., O'Sullivan, G.C., O'Riordan, M., Holmes, E., Wang, Y., and Marchesi, J.R. (2008). Culture-independent analysis of the gut microbiota in colorectal cancer and polyposis. Environmental Microbiology *10*, 789-798.
94. Hope, M.E., Hold, G.L., Kain, R., and El-Omar, E.M. (2005). Sporadic colorectal cancer-role of the commensal microbiota. FEMS Microbiology Letters. 2005 Mar 1;244(1):1-7.
95. Clement, K., Vaisse, C., Lahlou, N., Cabrol, S., Pelloux, V., Cassuto, D., Gourmelen, M., Dina, C., Chambaz, J., Lacorte, J.M., et al. (1998). A mutation

in the human leptin receptor gene causes obesity and pituitary dysfunction. Nature *392*, 398-401.
96. Bäckhed, F., Manchester, J.K., Semenkovich, C.F., and Gordon, J.I. (2007). Mechanisms underlying the resistance to diet-induced obesity in germ-free mice. Proceedings of the National Academy of Sciences *104*, 979-984.
97. Larsen, N., Vogensen, F.K., van den Berg, F.W.J., Nielsen, D.S., Andreasen, A.S., Pedersen, B.K., Al-Soud, W.A., Sørensen, S.J., Hansen, L.H., and Jakobsen, M. (2010). Gut Microbiota in Human Adults with Type 2 Diabetes Differs from Non-Diabetic Adults. PLoS ONE *5*, e9085.
98. Cani, P.D., Amar, J., Iglesias, M.A., Poggi, M., Knauf, C., Bastelica, D., Neyrinck, A.M., Fava, F., Tuohy, K.M., Chabo, C., et al. (2007). Metabolic Endotoxemia Initiates Obesity and Insulin Resistance. Diabetes *56*, 1761-1772.
99. Moore, W.E., and Holdeman, L.V. (1974). Human fecal flora: the normal flora of 20 Japanese-Hawaiians. Applied Microbiology. 1974 May;27(5):961-79.
100. Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. Genome Biology *3*.
101. Staley, J.T., and Konopka, A. (1985). Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. Annual Review of Microbiology *39*, 321-346.
102. Simpson, J.M., McCracken, V.J., White, B.A., Gaskins, H.R., and Mackie, R.I. (1999). Application of denaturant gradient gel electrophoresis for the analysis of the porcine gastrointestinal microbiota. J Microbiol Methods *36*, 167-179.
103. Dicksved, J., Halfvarson, J., Rosenquist, M., Jarnerot, G., Tysk, C., Apajalahti, J., Engstrand, L., and Jansson, J.K. (2008). Molecular analysis of the gut microbiota of identical twins with Crohn's disease. ISME J *2*, 716-727.
104. Marianne, H.F., Paul, W.O.T., and Eamonn, M.M.Q. (2012). Techniques used to characterize the gut microbiota: a guide for the clinician. Nature Reviews Gastroenterology and Hepatology *9*, 312-322.
105. Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., and Hugenholtz, P. (2008). A Bioinformatician's Guide to Metagenomics. Microbiology and Molecular Biology Reviews *72*, 557-578.
106. Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., and Goodman, R.M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & Biology *5*, R245-R249.
107. Pace, N.R., Stahl,D.A,. Lane,.L.L.,Olesn.G.J. (1986). The analysis of natural microbial populations by rRNA sequences. Advances in microbial ecology *9*, 1-55.
108. Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiology and Molecular Biology Reviews. *68*, 669-685.
109. Ekkers, D.M., Cretoiu, M.S., Kielak, A.M., and Elsas, J.D. (2011). The great screen anomaly-a new frontier in product discovery through functional metagenomics. Applied Microbiology and Biotechnology *93*, 1005-1020.
110. Cowan, D., Meyer, Q., Stafford, W., Muyanga, S., Cameron, R., and Wittwer, P. (2005). Metagenomic gene discovery: past, present and future. Trends in Biotechnology *23*, 321-329.
111. Bertrand, H.l.n., Poly, F., Van, V.T., Lombard, N., Nalin, R., Vogel, T.M., and Simonet, P. (2005). High molecular weight DNA recovery from soils prerequisite for biotechnological metagenomic library construction. Journal of Microbiological Methods *62*, 1-11.

112. Salonen, A., Nikkila, J., Jalanka-Tuovinen, J., Immonen, O., Rajilic-Stojanovic, M., Kekkonen, R.A., Palva, A., and de Vos, W.M. (2010). Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. Journal of Microbiological Methods *81*, 127-134.
113. Torres-Cortes, G., Millan, V., Ramirez-Saad, H.C., Nisa-Martinez, R., Toro, N., and Martinez-Abarca, F. (2011). Characterization of novel antibiotic resistance genes identified by functional metagenomics on soil samples. Environmental Microbiology *13*, 1101-1114.
114. Rondon, M.R., August, P.R., Bettermann, A.D., Brady, S.F., Grossman, T.H., Liles, M.R., Loiacono, K.A., Lynch, B.A., MacNeil, I.A., Minor, C., et al. (2000). Cloning the Soil Metagenome: a Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms. Applied and Environmental Microbiology *66*, 2541-2547.
115. Lepage, P., Leclerc, M.C., Joossens, M., Mondot, S., Blottiére, H.M., Raes, J., Ehrlich, D., and Doré, J.(2011). A metagenomic insight into our gut's microbiome. Gut.*62*, 146-148.
116. Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. Microbial Informatics and Experimentation *2*, 3-3.
117. Grada, A., and Weinbrecht, K. (2013). Next-Generation Sequencing: Methodology and Application. J Investigative Dermatology *133*, e11.
118. MacLean, D., Jones, J.D.G., and Studholme, D.J. (2009). Application of 'next-generation' sequencing technologies to microbial genetics. Nature Reviews Microbiology *7*, 287-296.
119. Chou, H.H., and Holmes, M.H. (2001). DNA sequence quality trimming and vector removal. Bioinformatics *17*, 1093-1104.
120. Lapidus, A.L. (2009). Genome Sequence Databases (Overview): Sequencing and Assembly.
121. Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Federhen, S., et al. (2011). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res *39*, 38-51.
122. Lukashin, A.V., and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res *26*, 1107-1115.
123. Noguchi, H., Park, J., and Takagi, T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Research *34*, 5623-5630.
124. Frishman, D., Mironov, A., Mewes, H.-W., and Gelfand, M. (1998). Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. Nucleic Acids Research *26*, 2941-2947.
125. Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. (1999). Improved microbial gene identification with GLIMMER. Nucleic Acids Res *27*, 4636-4641.
126. Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., et al. (2004). The Pfam protein families database. Nucleic Acids Research *40*, D290-D301.
127. Markowitz, V.M., Ivanova, N., Palaniappan, K., Szeto, E., Korzeniewski, F., Lykidis, A., Anderson, I., Mavromatis, K., Kunin, V., Garcia Martin, H., et al. (2006). An experimental metagenome data management and analysis system. Bioinformatics *22*, e359-367.

128. Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A genomic perspective on protein families. Science *278*, 631-637.
129. Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S., and Banfield, J.F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature *428*, 37-43.
130. Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., et al. (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science *304*, 66-74.
131. Beja, O., Aravind, L., Koonin, E.V., Suzuki, M.T., Hadd, A., Nguyen, L.P., Jovanovich, S.B., Gates, C.M., Feldman, R.A., Spudich, J.L., et al. (2000). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science *289*, 1902-1906.
132. Prakash, T., and Taylor, T.D. (2012). Functional assignment of metagenomic data: challenges and applications. Briefings in Bioinformatics *13*, 711-727.
133. Culligan, E.P., Sleator, R.D., Marchesi, J.R., and Hill, C. (2012). Functional metagenomics reveals novel salt tolerance loci from the human gut microbiome. ISME J *6*, 1916-1925.
134. Simon, C., Herath, J., Rockstroh, S., and Daniel, R. (2009). Rapid Identification of Genes Encoding DNA Polymerases by Function-Based Screening of Metagenomic Libraries Derived from Glacial Ice. Applied and Environmental Microbiology *75*, 2964-2968.
135. Rabausch, U., Juergensen, J., Ilmberger, N., Böhnke, S., Fischer, S., Schubach, B., Schulte, M., and Streit, W.R. (2013). Functional Screening of Metagenome and Genome Libraries for Detection of Novel Flavonoid-Modifying Enzymes. Applied and Environmental Microbiology.
136. Uchiyama, T., Abe, T., Ikemura, T., and Watanabe, K. (2005). Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. Nature Biotechnology *23*, 88-93.
137. Uchiyama, T., and Miyazaki, K.(2010). Product-Induced Gene Expression, a Product-Responsive Reporter Assay Used To Screen Metagenomic Libraries for Enzyme-Encoding Genes. Applied and Environmental Microbiology *76*, 7029-7035.
138. Williamson, L.L., Borlee, B.R., Schloss, P.D., Guan, C., Allen, H.K., and Handelsman, J. (2005). Intracellular Screen To Identify Metagenomic Clones That Induce or Inhibit a Quorum-Sensing Biosensor. Applied and Environmental Microbiology *71*, 6335-6344.
139. Tasse, L., Bercovici, J., Pizzut-Serin, S., Robe, P., Tap, J., Klopp, C., Cantarel, B.L., Coutinho, P.M., Henrissat, B., Leclerc, M., et al. (2010). Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. In Genome research, Volume 20. (United States), pp. 1605-1612.
140. Craig, J.W., Chang, F.-Y., Kim, J.H., Obiajulu, S.C., and Brady, S.F. Expanding Small-Molecule Functional Metagenomics through Parallel Screening of Broad-Host-Range Cosmid Environmental DNA Libraries in Diverse Proteobacteria. Applied and Environmental Microbiology *76*, 1633-1641.
141. Owen, J.G., Robins, K.J., Parachin, N.S., and Ackerley, D.F. (2012). A functional screen for recovery of 4'-phosphopantetheinyl transferase and associated natural product biosynthesis genes from metagenome libraries. Environmental Microbiology. 2012 May;14(5):1198-209. doi: 10.1111/j.1462-2920.2012.02699.x. Epub 2012 Feb 22.

142. Pang, H., Zhang, P., Duan, C.-J., Mo, X.-C., Tang, J.-L., and Feng, J.-X. (2009). Identification of Cellulase Genes from the Metagenomes of Compost Soils and Functional Characterization of One Novel Endoglucanase. Current Microbiology *58*, 404-408.
143. Nacke, H., Will, C., Herzog, S., Nowka, B., Engelhaupt, M., and Daniel, R.(2011). Identification of novel lipolytic genes and gene families by screening of metagenomic libraries derived from soil samples of the German Biodiversity Exploratories. FEMS Microbiology Ecology *78*, 188-201.
144. McGarvey, K.M., Queitsch, K., and Fields, S. (2012). Wide Variation in Antibiotic Resistance Proteins Identified by Functional Metagenomic Screening of a Soil DNA Library. Applied and Environmental Microbiology *78*, 1708-1714.
145. Jones, B.V., and Marchesi, J.R. (2007). Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome. Nat Meth *4*, 55-61.
146. Rui He, B.W., Toshiyuki Wakimoto, Manyuan Wang, Liancai Zhu and Ikuro Abe (2013). Cyclodipeptides from Metagenomic Library of a Japanese Marine Sponge. Journal of Brazilian biochemistry *24*, 1926-1932.
147. Lang, K.S., Anderson, J.M., Schwarz, S., Williamson, L., Handelsman, J., and Singer, R.S. (2010). Novel Florfenicol and Chloramphenicol Resistance Gene Discovered in Alaskan Soil by Using Functional Metagenomics. Applied and Environmental Microbiology *76*, 5321-5326.
148. Jeon, J.H., Kim, J.T., Lee, H.S., Kim, S.-J., Kang, S.G., Choi, S.H., and Lee, J.-H. (2011). Novel Lipolytic Enzymes Identified from Metagenomic Library of Deep-Sea Sediment. Evidence-Based Complementary and Alternative Medicine *2011*, 9.
149. Berlemont, R., Delsaute, M., Pipers, D., D'Amico, S., Feller, G., Galleni, M., and Power, P. (2009). Insights into bacterial cellulose biosynthesis by functional metagenomics on Antarctic soil samples. ISME J *3*, 1070-1081.
150. Cecchini, D.A., Laville, E., Laguerre, S., Robe, P., Leclerc, M., Doré, J.l., Henrissat, B., Remaud-Siméon, M., Monsan, P., and Potocki-Véronèse, G. (2013). Functional Metagenomics Reveals Novel Pathways of Prebiotic Breakdown by Human Gut Bacteria. PLoS ONE *8*, e72766.
151. Shao, H., Xu, L., and Yan, Y. (2013). Isolation and characterization of a thermostable esterase from a metagenomic library. J Industrial Microbiology and Biotechnology *40*, 1211-1222.
152. Sophie Biver, D.P.a.M.V. (2013). Characterization of a new oxidant-stable serine protease isolated by functional metagenomics. SpringerPlus *2*.
153. Cheng, G., Hu, Y., Yin, Y., Yang, X., Xiang, C., Wang, B., Chen, Y., Yang, F., Lei, F., Wu, N., et al. (2012). Functional screening of antibiotic resistance genes from human gut microbiota reveals a novel gene fusion. FEMS Microbiology Letters *336*, 11-16.
154. Lussier, F.-X., Chambenoit, O., Côté, A., Hupé, J.-F., Denis, F., Juteau, P., Beaudet, R., and Shareck, F. (2010). Construction and functional screening of a metagenomic library using a T7 RNA polymerase-based expression cosmid vector. Journal of Industrial Microbiology and Biotechnology, 1-8.
155. Neveu, J., Regeard, C., and DuBow, M. (2011). Isolation and characterization of two serine proteases from metagenomic libraries of the Gobi and Death Valley deserts. Applied Microbiology and Biotechnology *91*, 635-644.
156. Donato, J.J., Moe, L.A., Converse, B.J., Smart, K.D., Berklein, E.C., McManus, P.S., and Handelsman, J. (2010). Metagenomic analysis of apple orchard soil

reveals antibiotic resistance genes encoding predicted bifunctional proteins. Applied and Environmental Microbiology *76*, 4396-4401.
157. Yao, J., Fan, X.J., Lu, Y., and Liu, Y.H. (2011). Isolation and Characterization of a Novel Tannase from a Metagenomic Library. Journal of Agricultural and Food Chemistry *59*, 3812-3818.
158. Courtois, S., Cappellano, C.M., Ball, M., Francou, F.-X., Normand, P., Helynck, G., Martinez, A., Kolvek, S.J., Hopke, J., Osburne, M.S., et al. (2003). Recombinant Environmental Libraries Provide Access to Microbial Diversity for Drug Discovery from Natural Products. Appl. Environmental Microbiology. *69*, 49-55.
159. Thompson, S.A., Maani, E.V., Lindell, A.H., King, C.J., and McArthur, J.V. (2007). Novel Tetracycline Resistance Determinant Isolated from an Environmental Strain of Serratia marcescens. Applied and Environmental Microbiology. *73*, 2199-2206.
160. Turnbaugh, P.J., Henrissat, B., and Gordon, J.I. (2010). Viewing the human microbiome through three-dimensional glasses: integrating structural and functional studies to better define the properties of myriad carbohydrate-active enzymes. Acta Crystallogr Sect F Struct Biol Cryst Commun *66*, 1261-1264.
161. Ekkers, D., Cretoiu, M., Kielak, A., and Elsas, J. (2012). "The great screen anomaly"-a new frontier in product discovery through functional metagenomics. Applied Microbiology and Biotechnology *93*, 1005-1020.
162. Hall, T. (2005). BioEdit. Biological sequence alignment editor for Win95.
163. Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular Biology and Evolution.
164. Rawlings, N.D., Barrett, A.J., and Bateman, A. (2012). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Research *40*, D343-D350.
165. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., and Henrissat, B. (2013). The carbohydrate-active enzymes database (CAZy) in 2013. In Nucleic Acids Research, Volume 42. (England), pp. D490-495.
166. Fischer, M., and Pleiss, J. (2003). The Lipase Engineering Database: a navigation and analysis tool for protein families. Nucleic Acids Res *31*, 319-321.
167. Schomburg, D., and Schomburg, I. Enzyme databases. (2010). Methods in Molecular Biology *609*, 113-128.
168. Shen, H.-B., and Chou, K.-C. (2009). Identification of proteases and their types. Analytical Biochemistry *385*, 153-160.
169. D. M, B., J. J, B., and B. S, H. (1969). Role of a Buried Acid Group in the Mechanism of Action of Chymotrypsin. Nature *221*, 337-340.
170. Ebeling, W., Hennrich, N., Klockow, M., Metz, H., Orth, H.D., and Lang, H. (1974). Proteinase K from Tritirachium album Limber. European Journal of Biochemistry *47*, 91-97.
171. Hay, R.T. (2007). SUMO-specific proteases: a twist in the tail. Trends in Cellular Biology *17*, 370-376.
172. Sippl, W., Collura, V., and Colland, F. (2011). Ubiquitin-specific proteases as cancer drug targets. Future Oncology. 2011 May;7(5):619-32. doi: 10.2217/fon.11.39.
173. Yan, R., Bienkowski, M.J., Shuck, M.E., Miao, H., Tory, M.C., Pauley, A.M., Brashler, J.R., Stratman, N.C., Mathews, W.R., Buhl, A.E., et al. (1999).

Membrane-anchored aspartyl protease with Alzheimer's disease [beta]-secretase activity. Nature *402*, 533-537.
174. Koblinski, J.E., Ahram, M., and Sloane, B.F. (2000). Unraveling the role of proteases in cancer. Clinical Chimica Acta *291*, 113-135.
175. Brik, A., and Wong, C.H. (2003). HIV-1 protease: mechanism and drug discovery. Organic Biomolecular Chemistry. 2003 Jan 7;1(1):5-14.
176. Rao, M.B., Tanksale, A.M., Ghatge, M.S., and Deshpande, V.V. (1998). Molecular and Biotechnological Aspects of Microbial Proteases. Microbiology and Molecular Biology Reviews. *62*, 597-635.
177. Ten, L.N., Im, W.T., Kim, M.K., and Lee, S.T. (2005). A plate assay for simultaneous screening of polysaccharide- and protein-degrading micro-organisms. Letters in Applied Microbiology *40*, 92-98.
178. Wolff, A.M., Showell, M.S., Venegas, M.G., Barnett, B.L., and Wertz, W.C. (1996). Laundry performance of subtilisin proteases. Advanced Experimental Medical Biology. 1996;379:113-20.
179. Gupta, R., and Ramnani, P. (2006). Microbial keratinases and their prospective applications: an overview. Applied Microbiology and Biotechnology *70*, 21-33.
180. Riley, K.N., and Herman, I.M. (2005). Collagenase promotes the cellular responses to injury and wound healing in vivo. Journal of Burns and Wounds *4*.
181. Jordan, G.H. (2008). The use of intralesional clostridial collagenase injection therapy for Peyronie's disease: a prospective, single-center, non-placebo-controlled study. Journal of Sexual Medicine. 2008 Jan;5(1):180-7. doi: 10.1111/j.1743-6109.2007.00651.x.
182. Duval, M., Suciu, S., Ferster, A., Rialland, X., Nelken, B., Lutz, P., Benoit, Y., Robert, A., Manel, A.-M., Vilmer, E., et al. (2002). Comparison of *Escherichia coli* asparaginase with Erwinia asparaginase in the treatment of childhood lymphoid malignancies: results of a randomized European Organisation for Research and Treatment of Cancer Children's Leukemia Group phase 3 trial. Blood *99*, 2734-2739.
183. Lantz, M.S. (1997). Are bacterial proteases important virulence factors? Journal of Periodontal Research *32*, 126-132.
184. Maeda, H., and Molla, A. (1989). Pathogenic potentials of bacterial proteases. Clinica Chimica Acta *185*, 357-367.
185. Laarman, A.J., Ruyken, M., Malone, C.L., van Strijp, J.A., Horswill, A.R., and Rooijakkers, S.H. (2011). *Staphylococcus aureus* metalloprotease aureolysin cleaves complement C3 to mediate immune evasion. In Journal of immunology (Baltimore, Md. : 1950), Volume 186. (United States), pp. 6445-6453.
186. Young, J.A., and Collier, R.J. (2007). Anthrax toxin: receptor binding, internalization, pore formation, and translocation. Annual Reviews in Biochemisrty *76*, 243-265.
187. Williamson, L.C., Halpern, J.L., Montecucco, C., Brown, J.E., and Neale, E.A. (1996). Clostridial neurotoxins and substrate proteolysis in intact neurons: botulinum neurotoxin C acts on synaptosomal-associated protein of 25 kDa. Journal of Biological Chemistry. 1996 Mar 29;271(13):7694-9.
188. Hoy, B., Geppert, T., Boehm, M., Reisen, F., Plattner, P., Gadermaier, G., Sewald, N., Ferreira, F., Briza, P., Schneider, G., et al. (2012). Distinct roles of secreted HtrA proteases from gram-negative pathogens in cleaving the junctional protein and tumor suppressor E-cadherin. Journal of Biological Chemistry. 2012 Mar 23;287(13):10115-20. doi: 10.1074/jbc.C111.333419. Epub 2012 Feb 15.

189. Kuang, Z., Hao, Y., Walling, B.E., Jeffries, J.L., Ohman, D.E., and Lau, G.W.(2011). *Pseudomonas aeruginosa* elastase provides an escape from phagocytosis by degrading the pulmonary surfactant protein-A. PLoS One *6*.
190. Park, S.Y., Shin, Y.P., Kim, C.H., Park, H.J., Seong, Y.S., Kim, B.S., Seo, S.J., and Lee, I.H. (2008). Immune Evasion of *Enterococcus faecalis* by an Extracellular Gelatinase That Cleaves C3 and iC3b. The Journal of Immunology *181*, 6328-6336.
191. Loomes, L.M., Senior, B.W., and Kerr, M.A. (1990). A proteolytic enzyme secreted by *Proteus mirabilis* degrades immunoglobulins of the immunoglobulin A1 (IgA1), IgA2, and IgG isotypes. Infection and Immunity *58*, 1979-1985.
192. Holton, J. (2008). Enterotoxigenic *Bacteroides fragilis*. Current Infectious Disease Reports *10*, 99-104.
193. Grys, T.E., Walters, L.L., and Welch, R.A. (2006). Characterization of the StcE Protease Activity of *Escherichia coli* O157:H7. Journal of Bacteriology *188*, 4646-4653.
194. Harrington, D.J. (1996). Bacterial collagenases and collagen-degrading enzymes and their potential role in human disease. Infect Immun *64*, 1885-1891.
195. Szabady, R.L., Yanta, J.H., Halladin, D.K., Schofield, M.J., and Welch, R.A.(2011). TagA is a secreted protease of *Vibrio cholerae* that specifically cleaves mucin glycoproteins. Microbiology *157*, 516-525.
196. Dave, J.A., Gey van Pittius, N.C., Beyers, A.D., Ehlers, M.R.W., and Brown, G.D. (2002). Mycosin-1, a subtilisin-like serine protease of *Mycobacterium tuberculosis*, is cell wall-associated and expressed during infection of macrophages. BMC Microbiology *2*, 30-30.
197. Ohol, Y.M., Goetz, D.H., Chan, K., Shiloh, M.U., Craik, C.S.,and Cox, J.S.(2010). *Mycobacterium tuberculosis* MycP1 Protease Plays a Dual Role in Regulation of ESX1 Secretion and Virulence.Cell Host & Microbe *7*,210-220.
198. Macfarlane, G.T., Cummings, J.H., and Allison, C. (1986). Protein Degradation by Human Intestinal Bacteria. J Genetic Microbiology *132*, 1647-1656.
199. Chacko, A., and Cummings, J.H. (1988). Nitrogen losses from the human small bowel: obligatory losses and the effect of physical form of food. Gut *29*, 809-815.
200. Gibson, S.A., McFarlan, C., Hay, S., and MacFarlane, G.T. (1989). Significance of microflora in proteolysis in the colon. Appl Environ Microbiol *55*, 679-683.
201. Toivola, D.M., Krishnan, S., Binder, H.J., Singh, S.K., and Omary, M.B. (2004). Keratins modulate colonocyte electrolyte transport via protein mistargeting. J Cell Biol *164*, 911-921.
202. Steck, N., Hoffmann, M., Sava, I.G., Kim, S.C., Hahne, H., Tonkonogy, S.L., Mair, K., Krueger, D., Pruteanu, M., Shanahan, F., et al. (2011). *Enterococcus faecalis* metalloprotease compromises epithelial barrier and contributes to intestinal inflammation. Gastroenterology *141*, 959-971.
203. Carroll, I.M., and Maharshak, N. (2013). Enteric bacterial proteases in inflammatory bowel disease- pathophysiology and clinical implications. World Journal of Gastroenterology. *19*, 7531-7543.
204. Steck, N., Mueller, K., Schemann, M., and Haller, D. (2013). Republished: Bacterial proteases in IBD and IBS. Postgraduate Medical Journal *89*, 25-33.
205. Lorenz, P., and Eck, J. (2005). Metagenomics and industrial applications. Nature Reviews Microbiology *3*, 510-516.

206. Allen, H.K., Donato, J., Wang, H.H., Cloud-Hansen, K.A., Davies, J., and Handelsman, J. (2010). Call of the wild: antibiotic resistance genes in natural environments. Nature Reviews Microbiology *8*, 251-259.
207. Jones, B.V., Sun, F., and Marchesi, J.R. (2007). Using skimmed milk agar to functionally screen a gut metagenomic library for proteases may lead to false positives. Letters in Applied Microbiology *45*, 418-420.
208. Rondon, M.R., August, P.R., Bettermann, A.D., Brady, S.F., Grossman, T.H., Liles, M.R., Loiacono, K.A., Lynch, B.A., MacNeil, I.A., Minor, C., et al. (2000). Cloning the Soil Metagenome: a Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms. Applied Environmental Microbiology. *66*, 2541-2547.
209. Waschkowitz, T., Rockstroh, S., and Daniel, R. (2009). Isolation and Characterization of Metalloproteases with a Novel Domain Structure by Construction and Screening of Metagenomic Libraries. Applied Environmental Microbiology. *75*, 2506-2516.
210. Lopez-Otin, C., and Overall, C.M. (2002). Protease degradomics: A new challenge for proteomics. Nature Reviews Molecular and Cellular Biology *3*, 509-519.

2.0 GENERAL MATERIALS AND METHODS

Unless stated otherwise the following reagents were obtained from:

- PCR reagents were supplied New England Biolabs (NEB) UK
- Cloning reagents were supplied by ThermoFisher Scientific (Loughborough, UK)
- Media and chemicals were supplied by Sigma Aldrich (Dorest, UK)
- Primers were supplied by Sigma Aldrich (Dorset, UK).

2.1 Bacterial Strains and plasmids

The bacterial strains and plasmids used in this thesis, along with their growth requirements and the chapters to which they are relevant are indicated in Table 2.1.

Table 2.1 Bacterial strains, growth conditions and plasmids used in this thesis. *LFSMA*: Lactose free skimmed milk agar, *LB*: Luria-betani agar, *MRA*: de Man Rogosa and Sharpe agar *FM*: Functional metagenomic, *TSA*: Tryptone soy agar, *Amp*: Ampicillin, *Cm*:Chloramphenicol, *Km*: Kanamycin, *neo*: Neomycin

Species/ plasmid	Use	Media	Temp	Antibiotic	Ref/Source	Chapter
***B. subtilis* MY2016**	Positive control for protease activity	LFSMA	30	Amp (100µg/ml)	Marchesi Culture collection	3,5
***E. coli* MegaX DH10B™ T1^{R} Electrocompetent cells**	Host for metagenomic library	LB	37	Cm (12.5µg/ml)	Epicentre, Madison, WI, USA	5
Glycoside hydrolase clones	Testing LFSMA	LFSMA	37	Cm (12.5µg/ml)	[1]	3
L. bulgaricus	Testing LFSMA	MRS/LFSMA (anaerobic jar)	37	n/a	University College Cork culture collection	3
pGEMt®T-easy vector	Sub-cloning	n/a	n/a	Amp (100 µg/ml)	Promega, Madison, WI, USA	4
pNZ44	FM library construction	n/a	n/a	Cm (5 µg/ml)	[2]	5
S. thermophilus	Testing LFSMA	MRS/LFSMA (anaerobic jar)	37	n/a	University College Cork culture collection	3
XL1-blue super-competent subgrade *E. coli*	Subcloning and screening	LB	37	Amp (100 µg/ml)	Agilent, Berkshire, UK	4,5
***B. subtilis* WB800N**	Host for metagenomic library	TSA	30	Km or neo (10 µg/ml)	MoBiTec	5

pET-42a	Expression of M1-1	n/a	n/a	Km (10 μg/ml)	Marchesi Culture collection	4
***P. aeruginosa* PAO1**	To obtain *LasB* gene	TSA	37	n/a	Mahenthiralingam culture collection	5
***E. faecalis* (isolated in this work)**	To obtain gelE gene and positive control for TER and permeability assay	LFSMA	37	n/a	This work	5,7
***Burkholderia ambifaria* AMMD**	To test whether it forms halos on LF-SMA	LFSMA	37	n/a	Mahenthiralingam culture collection	4
***Burkholderia cenocepacia* PC184**	To test whether it forms halos on LF-SMA	LFSMA	37	n/a	Mahenthiralingam culture collection	4

2.1.1 Storage and maintenance of bacterial strains

All stocks of bacterial isolates were prepared by taking freshly grown isolates from their appropriate solid media which had been grown at appropriate temperatures (30°c or 37°C) overnight. Subsequent colony growth was re-suspended in appropriate liquid media containing 8% v/v DMSO and mixed to form a bacterial suspension prior to storage at -80°C.

2.2 Plasmids

Plasmids used in this thesis are also shown in Table 2.1

2.2.1 Plasmid Extraction

Plasmid and fosmid preparations from *E. coli* were carried out using the Qiaprep spin mini prep kit (QIAGEN LTD. West Sussex, UK) according to the manufacturer's protocol on cells from an overnight culture (no more than 16 hours at 37°C and shaking at 150 rpm) from a single colony from a fresh plate, in 5-10 mL of LB broth supplemented with the appropriate antibiotic. Following extraction, plasmids were viewed under UV light on a 0.8% (w/v) agarose gel supplemented with 1μg/ml ethidium bromide.

2.2.2 Storage and maintenance of plasmids

Bacteria used to maintain the plasmids after transformation were stored in the same manner as all other bacterial isolates with appropriate antibiotics.

2.3 Human Stool Samples

2.3.1 Sample collection

Institutional ethical approval was obtained prior to the study and was granted by Cardiff School of Biosciences, Cardiff University (Cardiff University's Research Ethics Committee). Participants gave written consent following retrieval of a participant information sheet detailing all aspects of the study and what their samples would be used for.

Faecal samples were collected from healthy male and female individuals (n=11) from Cardiff University and individuals suffering from IBD, male and female (n =13) were recruited as part of another study at St Mary's Hospital, Imperial College London. All samples were processed and stored following correct protocol as determined by the Human Tissue Act (HTA) 2004 and samples were all anonymised.

Samples were collected using disposable trays; participants were advised to ensure no urine was collected in the tray alongside the faecal sample. Trays were deposited anonymously in a fridge at 4 ˚C until DNA extraction (the same day). The remaining sample were thoroughly mixed and divided into 1 g samples in sterile universal tubes and immediately frozen at -20°C. An overview of the participants is provided in Table 2.2. Both male and female participants were analysed. The various medications for treatment of those with IBD are also listed. Patients that had not received antibiotic treatment in the last year were selected.

Table 2.2 shows available participant information including participant sex, diagnosis of IBD; CD:- Crohns disease, UC; Ulcerative colitis along with a harvery bradshaw activity index (HBAI) and Simple clinical colitis Activity index (SCCAI) value which are used to quantify the symptoms of Crohn's disease and ulcerative colitis respectively. A value of greater than or equal to 5 represents a clinically active disease. 'H' means healthy participant. Medications for IBD treatment and other ailments are also indicated. Following is a brief description of the purpose of each drug [a] immunosuppressive drug, [b] TNF inhibitor, [c] anti-inflammatory, [d] monoclonal antibody against TNF-α, [e] steroid, [f] calcium and vitamin d supplement, [g] antimuscarinic, [h] proton pump inhibitor, [I] painkiller, [j]anti-inflammatory, [k] hypolilidemic drug to lower cholesterol, [l] to control diabetes, [m] painkiller, [n] statin used to lower cholesterol, [o] probiotic, [p] treatment for herpes simplex virus, [q] treatment for inflammation of joints, [r] anti-inflammatory, [s] opoid receptor antagonist, [t] anti-histamine, [u] anti-inflammatory, [v] opoid painkiller, [w] supplement, [x] painkiller. The healthy participants had no long term medications and no significant medical history.

Sample Code	Gender	Diagnosis	HBAI	SCCAI	Medications	Medical history
IBD 1	F	CD	5		Azathiaprine[a], Adulimumab[b]	Nil
IBD 2	M	UC		0	Azathiaprine, Asacol[c]	Nil
IBD 12	F	CD	5		Azathiaprine, infliximab[d]	Nil
IBD 16	M	CD	18		Prednisolone[e], Azathiaprine, adcal[f], mebeverine[g], omeprazole[h]	Nil
IBD 18	F	CD	11		Adulimumab, analgesia[i]	Nil
IBD 19	F	UC	0	6	Prednisolone, pentasa[j], simvastatin[k]	hypercholesterolaemia
IBD 25	F	CD	0		Pentasa, insulin[l], aspirin[m], pravastatin[n]	Diabetes, hypercholesterolaemia
IBD 26	M	UC	0	5	Nil	Nil
IBD 27	F	CD	1		VSL3[o], Azathiaprine, aciclovir[p], sulphasalazine[q]	Osteoarthritis
IBD 28	M	UC		2	Asacol	Impaired glucose tolerance
IBD 29	M	UC		2	Predisolone, predisolone enemas, mesalazine[r], naltrexone[s]	Nil
IBD 30	F	CD	1		Loratidine[t], balsalazide[u], codeine phosphate[v], glucosamine[w]	Allergies/hayfever, osteoarthritis
IBD 31	M	UC		7	Azathiaprine, pentasa, omeprazole, paracetamol[x]	Allergic rhinitis
S1	M	H	n/a	n/a	nil	nil
S2	M	H	n/a	n/a	nil	nil
S3	M	H	n/a	n/a	nil	nil
S4	F	H	n/a	n/a	nil	nil
S5	F	H	n/a	n/a	nil	nil
S6	M	H	n/a	n/a	nil	nil
S7	F	H	n/a	n/a	nil	nil
S8	M	H	n/a	n/a	nil	nil
S9	M	H	n/a	n/a	nil	nil
S10	M	H	n/a	n/a	nil	nil

2.4 DNA Extractions

2.4.1 From bacterial cells

DNA was extracted from 1 ml of overnight bacterial culture grown in appropriate liquid media inoculated from a single colony grown from solid media. Cells were pelleted by centrifugation at 4000 x g for 2 min. The pellet was resuspended in 100 µLTE buffer (10mM Tris/HCL, pH8, 10mMEDTA, pH8) and transferred to a 2 mLscrew cap microcentrifuge tube containing 0.5 mL 0.1 mm diameter zirconium beads (Biospec Products, Bartlesville, Oklahoma) and 500 µL lysis buffer (50 mM Tris-HCL, pH8, 70 mM EDTA, pH8, 1% (w/v) SDS) with 20 µl 20 mg/ml proteinase k (Sigma Aldrich). Bacterial cells were lysed by a 30 sec pulse on the fast prep bead beater and incubation at 37°C for activity of proteinase K degradation of protein. Saturated Ammonium acetate (200 µl) was added followed by vigorous shaking by the vortex. Chloroform (600 µl) was added and vigorously mixed by vortex again. Tubes were centrifuged at 14,000 x g for 5 min. Clear supernatant, 700 µL, was transferred to a sterile 1.5 mL microcentrifuge tube followed by addition of isopropanol (700 µl) and thoroughly mixed by hand before incubation at -20°C for 30 min. Tubes were centrifuged at 14,000 x g for 10 min and the supernatant was removed taking care not to disturb the white pellet. The pellet was washed with ethanol (100 µL) and air dried in a laminar flow hood. The DNA was resuspended in 200 µLTE buffer.

2.4.2 From faeces

DNA was extracted from faecal samples as soon as possible. Samples were always kept cool prior to extraction to minimise degradation of DNA. DNA was extracted using the QIamp DNA stool kit (Qiagen) according to the manufacturer's protocol with slight modifications; an additional bead beating step was implemented following addition of the first buffer using 0.5 g 0.1 mm zirconia/silica beads (BioSpec products Inc, Bartlesville, OK 74005, USA) were added to the tubes and subject to 3 x 1 min bead beating using the FastPrep®-24 (MP Biomedicals, Solon, OH 44139, USA).

2.4.3 Chelex rapid preparation of DNA for colony PCR

A sterile toothpick was inserted into a freshly grown (overnight) bacterial colony, picked and resuspended in 100 µL5% chelex® (Biorad, Hertforshire, UK; autoclaved) by pipette rapidly and swirling. The sample was subject to 3 cycles of heating at 98°C for 5 min and cooling at 4°C for 5 min to extract the DNA. The sample was centrifuged briefly (>12,000 g) to pellet, cellular debris and the clear supernatant was removed as the DNA extract ready for subsequent use.

2.5 Restriction digests

2.5.1 Partial digest of genomic DNA

Partial digest of genomic or metagenomic DNA was carried out as follows; 20 µL 10X restriction enzyme buffer and 2 µL BSA (10 mg/mL) was added to the genomic or metagenomic DNA (200 µL, $\geq$ 0.1 µg/µl). A 40 µL aliquot was added to one tube and 20 µL aliquots were added to 9 further tubes. Tubes were kept on ice from this point. Restriction enzyme was added (10 units/ µL) to tube 1 and mixed by pipette aspiration. 20 µL of tube 1 was added to tube 2 and mixed and 20 µL of tube 2 was added to tube 3. This process was repeated up to tube 9 and tube 10 was with no restriction enzyme to serve as a negative (no-digest) control. Tubes were incubated for 1 hour by at the optimum temperature for restriction enzyme activity. Digested DNA was viewed under UV light on a 1.5% (w/v) agarose gel supplemented with 1µg/mL safeview DNA smears of appropriate size were extracted with a gel extraction kit (Qiagen) according to the manufacturer's protocol.

2.5.2 Restriction enzyme digestion of plasmid DNA

To plasmid DNA (up to 1 µg/mL), restriction enzyme buffer (2 µL) was added followed by addition of restriction enzyme (1µL). Deionised water was added to a total volume of 20 µL and the sample was incubated overnight at the optimum temperature for that particular restriction enzyme. Digested plasmids were visualised under UV light on a 0.8% agarose gel supplemented with safeview and compared against uncut plasmid. To purify the digested plasmid, samples were passed through a Microcon YM-100

centrifugal filter unit (Merck Millipore, Darmstadt, Germany) according to the manufacturer's protocol.

2.5.3 Polymerase chain reaction of 16S rRNA genes

PCR reaction for amplification of the 16S rRNA gene from bacteiral isolates and metagenomic DNA was conducted using the universial 16S rRNA gene primers:- 27F (5'- AGAGTTTGATCMTGGCTCAG-3') and 1492 R (5'-TACGGYTACCTTGTTACGACTT-3') [3]. Amplification was performed in a 25 µL reaction volume consisting of forward and reverse primers (0.2 µM of each), dNTPs (2 µM of each dNTP), 1X taq buffer (10 mM Tris-HCl, 50mM KCl, 1.5mM $MgCl_2$, pH 8.3 at 25°C),1x Bovine Serum Albumin (NEB), *Taq* DNA polymerase (1.25 units), template DNA (1 ng-1 µg good quality DNA) and nuclease-free water. PCR conditions were as follows; an initial denaturation step of 95 ˚C for 5 min followed by 30 cycles of heating at 94 ˚C for 30 s as a further denaturation step, 55 ˚C for 30 s for primer annealing and 72 ˚C for 2 min with a final extension step at 72 ˚C for 5 min. Following amplification, PCR product was examined by electrophoresis in a 1 % agarose gel with 1 µg/mL safeview and visulaised using UV light.

PCR products were purified using the Qiagen PCR purification kit according to their protocol.

2.6 Development of competence in bacterial isolates

2.6.1 Development of competence in *E. coli*

The *E. coli* cells used for transformations were purchased as chemically competent (XL1-Blue supercompetent cells) or electrocompetent cells (MegaX DH10B $T1^R$ Electrocompetent cells).

2.6.2 Development of competence in *B. subtilis* WB800N

For *B subtilis* WB800N, competence was achieved as follows; a single colony that had been growing on a TSA neo[10] plate, ideally for a few days, was used to inoculate 10 mL of MM competence medium (10 mL SMM medium: $(NH_4)_2SO_4$ 0.2 % w/v, K_2HPO_4 1.4 % w/v, KH_2PO_4 0.6 % w/v, $Na_3C_6H_5O_7$ 0.1 % w/v and $MgSO_4 \cdot 7H_2O$ 0.02 % w/v supplemented with 0.125 mL 40 % w/v glucose, 0.1 mL 2 mg/mL tryptophan solution, 0.06 mL 1M $MgSO_4$, 0.01 % 20 % w/v Casamino acids and 0.005 mL 0.22 % w/v Fe-NH_4 citrate in a 100 mL flask) overnight. The sizes of the vessel used for the culture were found to be critical for successful development of competence in order for enough aeration during growth. Overnight culture (0.6 mL) was used to inoculate 10 mL fresh MM competence medium which was incubated at 37 ˚C for 3 hours. From this point forward, it was critical to maintain the temperature of the cells at 37 ˚C. To develop competence, after this 3 hour incubation, 10 mL Starvation medium was added (10 mL SMM medium with 0.125 mL 40 % w/v glucose and 0.06 mL 1M $MgSO_4$) and incubation was continued for 2 hours at 37 ˚C.

2.7 Transformation procedures

2.7.1 Chemical Transformation of *E. coli*

XL1-Blue supercompetent (Agilent) were transformed according to the manufacturer's protocol.

2.7.2Chemical transformation of B. subtilis WB800N

B. subtilis WB800N was chemically transformed immediately following development of competence as described in section 2.6. Following 2 hours of incubation in starvation medium, 1-5 µL of plasmid DNA was mixed with 0.4 mL competent cells in 14 mL, pre-warmed (37 ˚C) sterile VWR falcon tubes (Leicestershire, UK). Cells were shaken for 45 min-1 hour before plating on selective medium (TSA supplemented with 5 µg/ mL Cm).

2.7.3 Electrotransformation procedure

MegaX DH10B T1^{R} Electrocompetent cells (Life Technologies, Paisley, UK) were transformed according to the manufacturer's protocol.

2.8 Activity based screening

2.8.1 Preparation of semi skimmed milk agar (SMA) and Lactose-free skimmed milk agar (LF-SMA)

SSMA was prepared as follows: 10% (w/v) semi-skimmed milk solution was prepared in deionised water and autoclaved at 121°C for 15 min. 1.5% (w/v) purified agar (Thermo scientific) was also prepared in deionised water and autoclaved at 121°C for 15 min. Upon sterilisation, both were kept at 55°C in a water bath, then the semi-skimmed milk solution was added to the agar to give a final volume of 1% (w/v) semi-skimmed milk.

Lactose-free semi-skimmed milk powder was obtained from Valio (suomikauppa.fi) and agar was prepared in the same manner as described above to give a final concentration of 1% (w/v). Purified agar could be replaced by any agar that supports the growth of a particular organism. Zones of clearing around an inoculated organism or protein extract indicated protease activity (LF-SMA) and β-galactosidase activity or protease activity (SMA).

2.8.2 Preparation of crude cell free extracts

In order to prepare crude cell free extracts (CCFE) for protease assays, cells were grown overnight in appropriate media supplemented with the appropriate antibiotic. An aliquot (1%) of this starting culture was used to inoculate 50 mL of fresh media and grown to an optical density of 0.5 (600nm). After this point cultures were left to grow for a further 3 hours. Cells were harvested by centrifugation at 4000 x g for 10 min. The pellet was re-suspended in 2 mL of phosphate buffered saline (PBS) and each sample was bead beaten (0.1 mm diameter glass beads, 0.5g) for 30 seconds and repeated a further 2 times with cooling on ice for 5 min between each beating. Samples were centrifuged at 20,000 x g for 10 min and the resulting supernatant was taken as the cell free extract.

2.9 Measuring protease activity

2.9.1 Azocasein assay

General protease activity was determined by measuring the release of acid-soluble substance from azocasein (Sigma-Aldrich) over a period of 3 hours after precipitation. Azocasein (5 mg/mL) was prepared in 50 mM Tris-HCL, pH 8. The protease samples (100 µL) were added to the azocasein solution (100 µL). The mixture was incubated at 37°C and the reaction was terminated by the addition of 400 µL of 10 % (w/v) trichloroacetic acid (TCA). Protein was precipitated by centrifugation at 12,000 x g for 5 min and the resulting supernatant was transferred to a clean tube containing 700 µL 525 mM NaOH. The absorbance was measured using a spectrophotometer at 442 nm. Each reaction was carried out in triplicate. Negative controls were prepared by setting up a reaction and immediately terminating the reaction with TCA. The resulting precipitate was taken as a negative control. To minimise background interference, a further negative control was set up with just water. Proteinase K (Sigma Aldrich) was used as a positive control.

In order to determine the optimum temperature for protease activity, the previously described assay was carried out, but incubated at the desired temperatures. The reactions were terminated and protein precipitated as described above. Similarly, to deduce pH optima, the pH of the azocasein solution was varied from 3 to 11 using Sodium acetate (pH range 3-5), Tris-HCL (pH range 6-8) and Glycine sodium hydroxide (pH range 9-11).

2.9.2 Azo-coll assay

Azo-coll, an azo-dye impregnated collagen was purchased from Sigma Aldrich and used to measure collagenase activity in faecal protein extracts. The assay was carried out as previously described [4] with the protease under examination being the faecal protein extract (1 mg/mL). Negative and positive controls were prepared as described in 2.9.1. Where TCA wasn't used, samples were just immediately measured on the spectrophotometer.

2.9.3 Keratin azure assay

Keratin azure was used to measure keratinase activity of the faecal protein extracts. Keratin azure (44% w/v) was purchased from Sigma Aldrich and was prepared in 50 mM Sodium Phosphate buffer at pH 7.5. Protease extract was added and the solution was incubated at 37°C for 3 hours. Samples were centrifuged at 12,000 x g for 10 min and absorbance was measured at 595nm. Negative and positive controls were prepared as previously described.

2.10 Protein extraction

2.10.1 Total and extracellular protein extract from faecal samples

Faecal sample (1 g) was thawed and a 10 % faecal slurry was prepared using sterile PBS by thorough mixing on a Genies Whirlimixer™. To prepare the crude total protein extract, the faecal slurry was divided into 2 mL RNase and DNase free lysing matrix tubes (MP Biomedicals) containing 1.4 mm ceramic spheres, 0.1 mm silica spheres and one 4 mm glass sphere. Samples were kept on ice throughout. The samples were subject to bead beating using a FastPrep-24 bead beater (MP Biomedicals) at a speed of (6.0 m/s) for 30 s with a period of 5 min on ice between each beating. Following beating, samples were centrifuged at 20,000 x g for 30 mins at 4 °C. The supernatants from the bead beaten tubes were filtered through a 30 µm filter tube (Whatman, 113V, GE Healthcare Life Sciences, Buckinghamshire, UK). The filtrate was taken as the total protein extract. For the extracellular only samples, this centrifugation step was conducted immediately instead of the bead beating step. Supernatant, after filtration, was transferred to new sterile tubes and taken as the crude protein extract. NaN_3 was added aseptically to each sample to a final concentration of 0.05% w/v neat samples, 10-fold and 100-fold dilution were used to estimate protein concentration using the bicinchononic method (BCA) according to the manufacturer's instructions (PIERCE, Rockford, IL, USA) and samples were normalised to1 mg/mL protein. Samples were kept on ice at all times and subsequent protease assays were conducted within 24 hours.

2.11 SDS PAGE

SDS-PAGE was performed using the Laemelli methods with a 12% polyacrylamide gel. Gels were stained with InstantBLue™ Coomassie based protein staining solution (Fermentas, UK).

2.12 Statistical Analysis

Results were usually represented as the mean ± S.E.M. Statistical significance between means of data was determined by the student's T test or ANOVA if data was parametric. For non-parametric data, the Mann Whitney-U test was used instead. All Correlation analysis was conducted to determine the Pearson correlation coefficient (parametric data) or the Spearman rank correlation coefficient (non-parametric data) of two variables. All analyses were conducted in R software [5].

2.13 Cell Culture

2.13.1 Handling of frozen cell suspension and Preparation of initial low density cell stock

Frozen vials of the human colon epithelial HT-29 cell line were initially obtained from Sigma (Sigma-Aldrich) and stored in the vapour phase of liquid nitrogen. For preparation of initial cell stock, a vial was rapidly thawed (approximately 2 min) with gentle agitation in a 37 °C water bath while keeping the 'O-ring' and cap out of the water to reduce the possibility of contamination. The vial was sterilised decontaminated by spraying with 70% ethanol and transferred to a BSL-2 aseptic laminar flow hood where all proceeding operations were carried out. The vial's contents were transferred to a centrifuge tube containing 10 mL complete culture medium which comprised; McCoy's 5a medium supplemented with 10% fetal bovine serum (FBS) and 5% penicillin-streptomycin solution (ATCC®, Middlesex, UK) and centrifuged at approximated 125 x g for 8-10 min. The resultant pellet was resuspended in complete growth medium at a ratio of and 1 mL cell suspension was removed for cell count and viability performed using a haemocytometer. The cells were dispensed into a T-25cm^2 culture flask (Corning Life Sciences, Amsterdam, The Netherlands) at a cell density of

approximately 1.6 x 10^5 cells/cm^2. Cells were routinely maintained at 37 °C in 5% CO_2 atmosphere in a suitable incubator. Cells were observed daily under an inverted microscope (make/model) to ensure the culture was not contaminated and were sub-cultured upon reaching 50% confluence (approximately 5.5 x 10^4 cells).

2.3.2 Cell line maintenance and Sub-culturing Procedure

After growing the cells overnight the medium was removed from the culture flask and cells were washed with 2 mL Dulbecco's phosphate buffered saline (Life Technologies) and 2 mL 0.25% (w/v) Trypsin-0.53 mM EDTA solution (ATCC®) was added quickly after washing. Cells were observed under an inverted microscope for cell layer dispersion, if cells were difficult to detach, flasks were incubated at 37 °C for 5-15 min. Fresh complete growth medium (6 to 8 mL) was added to the flasks and cells were aspirated with gently pipetting. Appropriate aliquots of the cell suspension (using a subcultivation ratio of 1:10 to 1:12) were added to new culture vessels and incubated at 37 °C in 5% CO_2 atmosphere.

2.3.3 Cryopreservation

Cells were harvested as described in sub-culturing procedure. Prior to cell detachment, the cryopreservation medium was prepared as follows; complete growth medium was supplemented with 5% (v/v) DMSO and kept on ice until use. Cell density and viability was recorded and the cells were centrifuged at 125 x g for 5 min at room temperature. The media was carefully removed by using an automatic pipette to leave the cell pellet. Cryopreservation media was added so as to achieve a cell density of between 3.0 X 10^6 to 4.0 X 10^6 cells/mL and cells were resuspended with gentle pipetting. Cells were dispensed in 1 mL aliquots into cryovials (Thermo-Scientific) and equilibrated at 4 °C for a maximum time of 10 min.

The day prior to cell cryopreservation, 250 mL isopropanol (Sigma-Aldrich) was added to a "Mr. Frosty" freezing container (Sigma-Aldrich) and placed at 4 °C overnight. On the day of cryopreservation, after equilibration at 4 °C cryovials were inserted into the slots in the "Mr Frosty" and transferred to a -80 °C freezer and left overnight. The next day the cryovials were transferred to the vapour phase of a liquid nitrogen freezer.

2.4 REFERENCES

1. Jones, B.V., Sun, F., and Marchesi, J.R. (2007). Using skimmed milk agar to functionally screen a gut metagenomic library for proteases may lead to false positives. Letters in Applied Microbiology *45*, 418-420.
2. McGrath, S., Fitzgerald, G.F., and van Sinderen, D. (2001). Improvement and optimization of two engineered phage resistance mechanisms in *Lactococcus lactis*. Appl Environ Microbiol. 2001 Feb;67(2):608-16.
3. Lane, D.J. (1991). 16S/23S rRNA sequencing. Nucleic Acid Techniques in Bacterial Systematics, 125-175.
4. Jiang, N., Tan, N.S., Ho, B., and Ding, J.L. (2007). Azocoll protease activity assay.
5. Team, R.C. (2012). R: A language and environment for statistical computing.

3.0 A ROBUST PLATE ASSAY FOR DETECTION OF EXTRACELLULAR MICROBIAL PROTEASE ACTIVITY IN METAGENOMIC SCREENS AND PURE CULTURES

The analyses and discussion within this chapter has been published in; [1] (*Morris, L.S., Evans, J., and Marchesi, J.R. (2012) A robust plate assay for detection of extracellular microbial protease activity in metagenomic screens and pure cultures. Journal of Microbiological Methods 91, 144-146.*)

3.1 INTRODUCTION

It is estimated that we are currently only able to cultivate as little as 1% of microorganisms from the any diverse environments of which they inhabit [2] . This realisation that there is a deficiency in current culture techniques to isolate most microorganisms from most environments has led to the emergence and progression of cultivation-independent techniques specifically '-omic' sequencing based technologies to explore the composition, diversity, structure, activity and genetic makeup of microbial communities [3]. The use of shotgun metagenomic sequencing and genomic sequencing (explained in full in chapter 1.0) has become a baseline technology for microbial community characterisation; however there are a number of limitations associated with the technique, particularly with regards to unearthing novel genetic data. Firstly, sequencing metagenomics is reliant on comparison of generated data to previously annotated sequences that are publicly available, should the generated sequence share low homology to any annotated sequence data, novel genes are likely to often be disregarded as ambiguous data. Unknown data can sometimes account for up to 70% of sequence generated from sequencing projects [4, 5] and while current resources available to annotate sequence data are progressing, they remain insufficient to cope with the enormous amounts of data being generated by the increasing number of metagenomic sequencing projects being conducted worldwide [5]. Secondly, rarely can an identified gene be comprehensively annotated in terms of their biochemical and functional capacity due to the fact that entire genes sequences are not often identified. Even if samples share significant homology with known genes then their function is still only putative and remains to be proven, validation of function of identified genes has to be conducted with further experimental analysis mainly involving expression of the

newly identified gene. Gene targeting methods have been implemented in order to circumvent this ambiguity. This process involves primers or probes that are designed following multiple alignments of conserved regions of sequences that encode a particular gene of interest [6]. There primers are used to screen and amplify metagenomic DNA samples for this gene of interest and further genomic walking techniques may be implemented in order to capture the entire functional, and potentially novel gene. This genome mining techniques has proved useful, but clearly a substantial amount of *priori* knowledge is required and thus rarely are novel, non-homologous genes identified.

To counteract the limitations of shotgun metagenome sequencing and directed PCR-amplification for the purposes of identifying novel functional genes, a form of phenotypic screening of metagenomic DNA is required to identify novel genes and characterize their functional capacity. Functional metagenomics (FM) provides an alternative, function-driven means for unearthing the vast and relatively untapped biological resources of the microbial world. Its' ability to do so has been demonstrated with the discovery of novel and biotechnologically useful molecules [7] [8] and clinically relevant antibiotic resistance genes [9]. FM is a methodological strategy to perform such metagenomic DNA screening, and relies on screening clones harbouring random fragments of metagenomic DNA (mgDNA). Clones are transformed or transfected with plasmids containing mgDNA. Individual clones are screened for chosen functional capabilities; this screening relies on the heterologous gene expression of the randomly inserted DNA and a method for high-throughput screening of clones

For a successful functional metagenomic screen, an observable and measurable phenotype to identify the expression and secretion of novel bio-products is needed, and research in to FM is limited by the ease and efficiency of the system used to identify the product of interest. Screening approaches for FM libraries range from the relatively 'low-tech' methods previously used to study and characterise the functions of pure cultures to the development of novel, high-throughput sophistacted FM library screening approaches. 'Low-tech' approaches include the identification of phenotypic alterations in factors such as colony morphology, colony pigmentation, motility etc. and then the use of chemical dye incorporated into agar as indicator medium such as the incorporation of BCIP (5-Bromo-4-chloro-3-indolyl phosphate) into a growth media

allows for the detection of alkaline phosphatase activity. Many enzymes require more sensitive means for identification such as High Performance Liquid Chromotography (HPLC) or the development of chromogenic or fluorescent substrate [10]. Research, particularly in the biotechnology field, is progressing towards the development of novel, high-throughput techniques for FM library screening Rabausch and colleagues developed a system known as metagenome extract thin-layer chromatography analysis (META) which allows the detection of glycosyltransferases [10]. Williamson and colleagues developed the METREX system, a novel intracellular system used to screen metagenomic clones involved with inserts associated with quorum sensing. Here, the host contained a biosensor for quorum-inducing molecules. Upon sensing, the cell produces GFP which is detected by fluorescent microscopy or fluorescence-based cell sorting. This suggests that similar methods could be implemented for unearthing other biologically active small molecules in metagenomic libraries [11]. Similarly, a substrate-induced gene expression (SIGEX) system was also developed. With this method, an operon-trap GFP expression vector is used to construct a metagenomic library. The entire library is subjected to a substrate-specific gene induction. Upon expression of such substrate-specific genes, the GFP is also expressed. High-throughput detection of GFP and consequently substrate-specific genes is carried out with fluorescence-activated cell sorting (FACS), allowing for detection of novel catabolic operons [12]. Uchiyama et al. also developed Product-induced gene expression (PIGEX) which shares the concepts of SIGEX except relies on the enzyme product and are unresponsive to the substrate. Here, the GFP reporter gene was downstream of the BenR gene; a gene encoding benzoate transcriptional precursor. However, the number of enzymes identified using what are considered the relatively ‘low-tech’ methodologies and the potential they have had both in unearthing novel functions of bacteria from diverse habitats and also in industry outweighs the discoveries made from the more complex techniques indicating that there is still a place for simple, robust and efficient screening means.

The function of particular interest in this body of research is microbial proteases as these are an important target for metagenomic screening due to their extensive possibility for use in industry [13] and also mainly due to their role as virulence factors [14-16]and putative role as initiators of inflammation of the gut epithelia [17-19]. However there is a current deficit in metagenomic screens successfully isolating these enzymes. While there are a number of reasons for this scarcity (discussed further in

chapter 5), one problem that we aim to resolve with this research is the current issue of false positives arising from the use of skimmed milk agar as a standard means for screening metagenomic libraries for proteolytic activity. Previous research by Jones *et al.* [20] found that an initial 231 metagenomic clones deemed positive for protease activity by the formation of distinctive halos of clearing around colonies on standard skimmed milk agar (SMA) were in fact, glycoside hydrolases which produced acetic acid from lactose fermentation and the pH drop was responsible for the phenotype, not protease activity. The demonstrated ineffectiveness of this agar has lead us to question the validity of putative proteases detected using this method, for example *Pailin et al*., [21] used SMA to identify extracellular protease activity in strains of the lactic acid bacteria; *Streptococcus thermophilus* and *Lactobacillus bulgaricus.*

AIMS

It can be concluded that while SMA could be appropriate for identifying lactose utilisation and the presence of galactosidase enzymes, it is not a sufficiently robust a screening technique for protease activity and an alternative needs to be developed in order to save time and money and to ultimately be an effective screening means for functional metagenomic screening. The aim of this research was to develop a simple and easy to prepare media as the substrate for detecting protease activity to allow discrimination against acid production.
Thus, the main aims of this research are:-

1. To develop a simple and robust media for the detection of extracellular proteases. We hypothesise that a milk-based media free of lactose (LFSMA) and other compounds that may interfere with a protease screen will be a much more robust means for screening for protease activity.

2. To utilise the developed media and test microorganisms previously deemed positive for extracellular protease activity.

3.2 RESULTS

3.2 .1 Agar plate assay

Skimmed milk agar was compared with a new version of milk-based agar that has been developed utilising lactose-free milk powder based on the hypothesis that many putatively positive protease producing isolates were in fact glycoside hydrolases and the positive phenotype observed was as a result of lactose fermentation as opposed to protein degradation.

It was found that the strains used were capable of degrading skimmed milk agar but not lactose-free milk agar. Similarly the metagenomic clones had a corresponding outcome whereas for the bacterial strains known to be proteolytically active; a positive phenotype was observed on both types of media (Figure. 3.1).

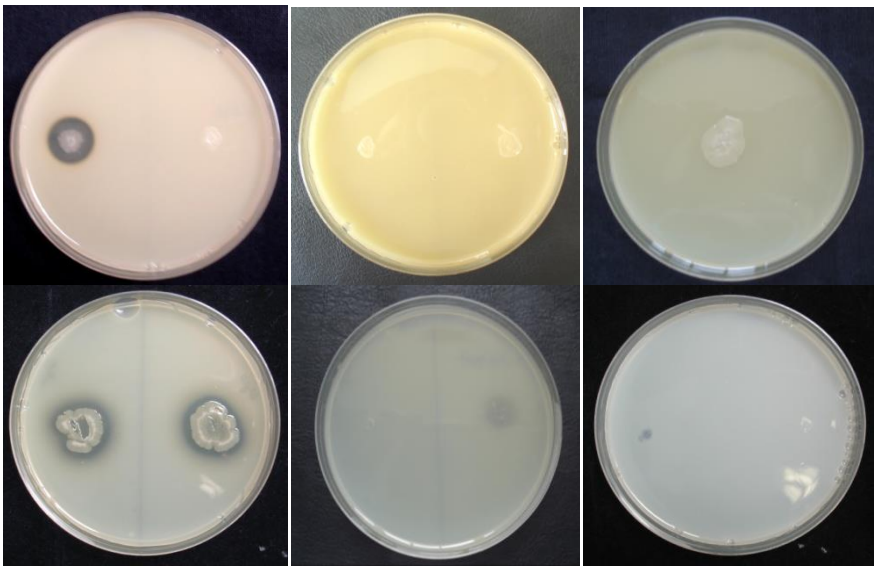


Figure 3.1. Plate assays for protease activity. From top left; *B. subtilis* MY2016 clearing LFSMA and a negative control, Top middle; *S. thermophilus* 859 and *L. bulgaricus*2483 failing to clear LFSMA, Top right; a glycoside hydrolase positive metagenomic clone failing to clear LFSMA, bottom left; two glycoside hydrolase positive metagenomic clones clearing SMA, bottom middle; *S. thermophilus* 859 clearing SMA, bottom right; *L. bulgaricus* 2483 beginning to clear SMA.

3.3.2 Azo-casein assay

For further confirmation that the proteolytic phenotype observed on SMA was in fact as a result of glycoside hydrolase activity and not proteases, cultures were further subject to azocasein assay (Figure. 3.2).*B. subtilis* MY2016 was shown to degrade the azocasein substrate at an average relative activity of 74% and glycoside hydrolase metagenomic clones showed less than 20% activity.

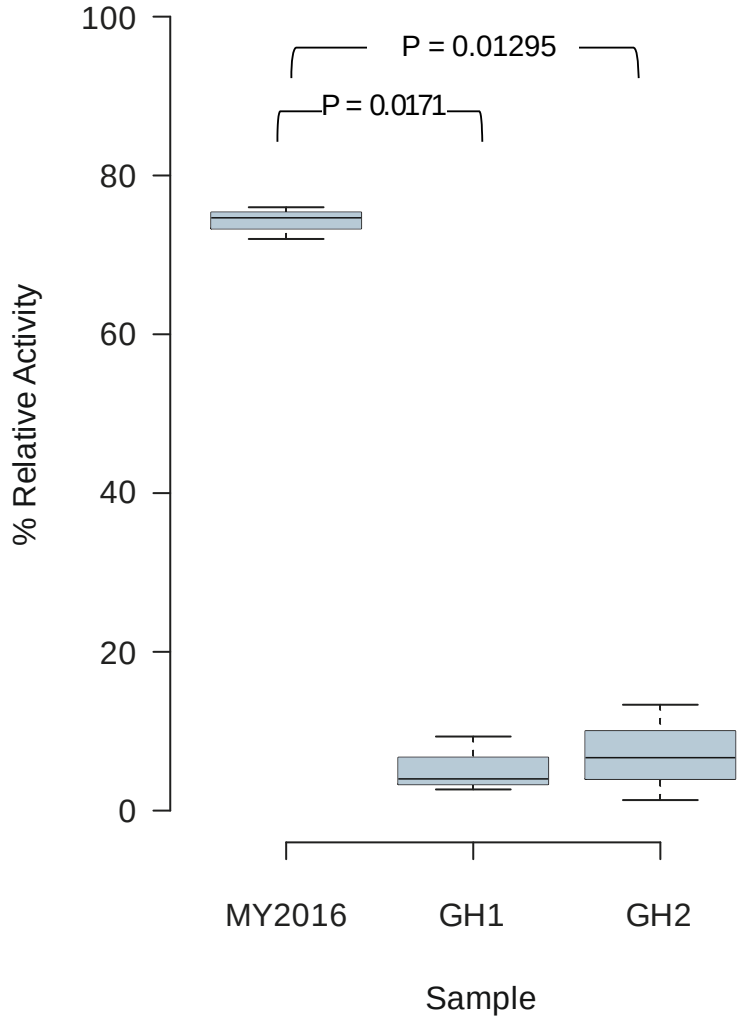


Figure 3.2. Protease activity as measured by the release of acid-soluble substance from azocasein (5 mg/ml) in HCl buffered Tris to a pH of 8.0 incubated at 37° C for 4 hours. Results are shown as the mean value of the results that were in triplicate and are shown as a relative percentage of total protease activity Centre lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles. 1= *B. subtilis* MY2016, 2 = Glycoside hydrolase clone, 3= glycoside hydrolase clone.

3.3 DISCUSSION

While functional metagenomics is clearly a promising technique for identifying novel microbial compounds, the process still faces a number of issues. Any steps towards mitigating these issues can only be seen as a positive for the future of metagenomics. In this study, we review the efficacy of standard skimmed milk agar as a means for identifying extracellular proteases since previous research has suggested skimmed milk is not a sufficiently robust screening means [20].

We compared this agar with a new version of milk-based agar that we have developed utilising lactose-free milk powder based on the hypothesis that many putatively positive protease producing isolates were in fact glycoside hydrolases and the positive phenotype observed was as a result of lactose fermentation as opposed to protein degradation.

It was found that the strains used were capable of degrading skimmed milk agar but not lactose-free milk agar. Similarly the metagenomic clones had a corresponding outcome whereas for the bacterial strains known to be proteolytically active; a positive phenotype was observed on both types of media (Figure. 3.1).

From our study, Valio™ lactose-free milk agar was found to be an effective and robust agar for correctly identifying proteases by way of distinct zones of clearing around a bacterial colony. From the screens undertaken in this study no false positives arose with the use of this type of agar. We would suggest that this agar would be much more appropriate for future screening of metagenomic libraries for protease activity. The robustness of this agar will allow for more efficient characterisation of enzyme activity which can then lead to identification of clinically or industrially relevant proteases.

3.4 MATERIALS AND METHODS

3.4.1 Strains and culture conditions

Bacillus subtilis MY2016 was used as a positive control for protease activity on all types of agar used in this study since it is known to secrete a number of extracellular proteases. Theses cultures were routinely grown on all agar at 30°C.
Strains of *Streptococcus thermophilus* 2483 and *Lactobacillus bulgaricus* 859 were obtained from the culture collection of University College cork, Department of Microbiology. Cultures were revived on MRS agar (Thermo Scientific, Oxoid) and incubated at 37°C in an anaerobic chamber (Merck, Darmstadt, Germany) and examined after 48 hour incubation.
False protease positive gylcoside hydrolase metagenomic clones were obtained from a previous metagenomic library [20] and were routinely revived from freezer stock on LB (Luria-bertani) agar supplemented with 12.5 μg/ml chloramphenicol.

3.4.2 Preparation of media used in this study

Agar plates (skimmed milk agar and valio ™ lactose free skimmed milk agar were prepared as described in the general methods section (Chapter 2.0)

3.4.3 Azo-casein assay to confirm protease activity

Azocasein assays were performed as described in the general methods section (Chapter 2.0).

3.4.4 Statistical analysis

Box plots of the relative percentage of protease activity determined following azocasein assay were generated using BoxPlotR [22] and the RColorBrewer package [23].
To determine statistical significance between the triplicate measurement of acid soluble substance released from the azocasein assay. A trypsin standard was used and incubated for the same period of time for generation of relative activity data. Normality tests were determined using R Software [24]. Statistical significance was determined using the Mann Whitney U test in R software also.

3.5 REFERENCES

1. Morris, L.S., Evans, J., and Marchesi, J.R. (2012). A robust plate assay for detection of extracellular microbial protease activity in metagenomic screens and pure cultures. Journal of Microbiological Methods *91*, 144-146.
2. Amann, R.I., Ludwig, W., and Schleifer, K.H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiological Reviews *59*, 143-169.
3. Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiology Molecular Biology Reviews. *68*, 669-685.
4. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59-65.
5. Prakash, T., and Taylor, T.D. (2012). Functional assignment of metagenomic data: challenges and applications. Briefings in Bioinformatics *13*, 711-727.
6. Iwai, S., Chai, B., Sul, W.J., Cole, J.R., Hashsham, S.A., and Tiedje, J.M. (2009). Gene-targeted-metagenomics reveals extensive diversity of aromatic dioxygenase genes in the environment. ISME J *4*, 279-285.
7. Uchiyama, T., and Miyazaki, K. (2009). Functional metagenomics for enzyme discovery: challenges to efficient screening. Current Opinion in Biotechnology *20*, 616-622.
8. Schloss, P.D., and Handelsman, J. (2003). Biotechnological prospects from metagenomics. Current Opinion in Biotechnology *14*, 303-310.
9. Allen, H.K., Donato, J., Wang, H.H., Cloud-Hansen, K.A., Davies, J., and Handelsman, J.(2010). Call of the wild: antibiotic resistance genes in natural environments. Nature Reviews Microbiology *8*, 251-259.
10. Rabausch, U., Juergensen, J., Ilmberger, N., Böhnke, S., Fischer, S., Schubach, B., Schulte, M., and Streit, W.R. (2013). Functional Screening of Metagenome and Genome Libraries for Detection of Novel Flavonoid-Modifying Enzymes. Applied and Environmental Microbiology.
11. Williamson, L.L., Borlee, B.R., Schloss, P.D., Guan, C., Allen, H.K., and Handelsman, J. (2005). Intracellular Screen To Identify Metagenomic Clones That Induce or Inhibit a Quorum-Sensing Biosensor. Applied and Environmental Microbiology *71*, 6335-6344.
12. Uchiyama, T., Abe, T., Ikemura, T., and Watanabe, K. (2005). Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. Nature Biotechnology *23*, 88-93.
13. Rao, M.B., Tanksale, A.M., Ghatge, M.S., and Deshpande, V.V. (1998). Molecular and Biotechnological Aspects of Microbial Proteases. Microbiology and Molecular Biology Reviews. *62*, 597-635.
14. Lantz, M.S. (1997). Are bacterial proteases important virulence factors? Journal of Periodontal Research *32*, 126-132.
15. Maeda, H., and Molla, A. (1989). Pathogenic potentials of bacterial proteases. Clinica Chimica Acta *185*, 357-367.
16. Miyoshi, S.-i., and Shinoda, S. (2000). Microbial metalloproteases and pathogenesis. Microbes and Infection *2*, 91-98.
17. Carroll, I.M., and Maharshak, N. (2013). Enteric bacterial proteases in inflammatory bowel disease- pathophysiology and clinical implications. World Journal of Gastroenterology. *19*, 7531-7543.

18. Steck, N., Mueller, K., Schemann, M., and Haller, D. (2013). Republished: Bacterial proteases in IBD and IBS. Postgraduate Medical Journal *89*, 25-33.
19. Steck, N., Hoffmann, M., Sava, I.G., Kim, S.C., Hahne, H., Tonkonogy, S.L., Mair, K., Krueger, D., Pruteanu, M., Shanahan, F., et al. (2011). Enterococcus faecalis metalloprotease compromises epithelial barrier and contributes to intestinal inflammation. Gastroenterology *141*, 959-971.
20. Jones, B.V., Sun, F., and Marchesi, J.R. (2007). Using skimmed milk agar to functionally screen a gut metagenomic library for proteases may lead to false positives. Letters in Applied Microbiology *45*, 418-420.
21. Pailin, T., Kang, D.H., Schmidt, K., and Fung, D.Y. (2001). Detection of extracellular bound proteinase in EPS-producing lactic acid bacteria cultures on skim milk agar. Letters in Applied Microbiology. 2001 Jul;33(1):45-9.
22. Spitzer, M., Wildenhain, J., Rappsilber, J., and Tyers, M. (2014). BoxPlotR: a web tool for generation of box plots. Nature Methods *11*, 121-122.
23. Neuwirth, E. (2007). RColorBrewer: ColorBrewer palettes. R package version 1.0-2.
24. Team, R.C. (2012). R: A language and environment for statistical computing.

4.0 SCREENING, ISOLATION AND CHARACTERISATION OF TWO FUNCTIONAL METAGENOME-DERIVED METALLOPROTEASES ISOLATED FROM A MILK WASTE TREATMENT PLANT

The analyses and discussion within this chapter has been published in; [47] (*Morris, L.S.and Marchesi, J.R. Current functional metagenomic approaches only expand the existing protease sequence space, but does not presently add any novelty to it. Current Microbiology 284.)*

4.1 INTRODUCTION

Proteases are degradative enzymes responsible for catalysing the hydrolysis of the peptide bonds between sequences of amino acids. Though ubiquitous and essential in all living organisms, those produced by the microbial world often take a primary position in our interest due to their relative ease of genetic manipulation, physiological and biochemical traits and genetic diversity [2] making them important and cost effective biocatalysts in the mass production of many commercial products such as food, detergents, textiles and pharmaceuticals [2, 3]. In addition to their application in industry, microbial proteases are also clinically relevant as proteases are thought to account for up to 5% of the genome [4] of infectious organisms and are often associated with the pathology of certain human diseases acting as significant virulence factors aiding the onset of infection [4, 5]. Recent research has suggested this number may be much higher [6]. The advance of culture-independent techniques for studying microbial communities may result in the isolation of novel proteases, potentially in many in microorganisms that we are currently unable to culture. Hence, culture-independent techniques are to be implemented if we are to recognize the true extent and capability of microbial proteases and from there, understand their physiological roles in the environments they inhabit. Culture-independent techniques, namely functional metagenomics (FM) are currently the only methodologies capable of unearthing novel genes. FM has aided the discovery of a plethora of novel compounds (please see table 5.1 chapter 1.0) from diverse ecosystems from soil to marine sponges to the human gut. The technique has also allowed us to elucidate how bacterial communities have adapted to their particular environment [7]. However, as discussed further in the succeeding

chapter, successful FM projects are limited due to the reliance of heterologous expression and secretion of the protein encoded by the inserted foreign DNA. Consequently, positive 'hits' (the frequency of detecting a positive clone) remain fairly low. When embarking on a project that aims to isolate proteolytic enzymes, this frequency is even lower for reasons described in depth in the proceeding chapter, but in brief; surrogate hosts may not have the cellular machinery required to faithfully transcribe and translate a particular proteases gene, proteases of the host organism may recognise and degrade foreign (inserted) protein, there is also a likelihood that the active product exerts a toxic effect on the surrogate host. Until recently there was also insufficient screening means for proteases [8], but our recent research has proven that the use of lactose free skimmed milk agar is a robust substrate for screening for proteases and minimises false positives [1](Chapter 3.0). Several research groups have been successful in their quest to unearth proteases from various bacterial metagenomes [9-16]. However, firstly the study by Purohit *et al.,* created a metagenomic library and uncovered a protease using degenerate PCR with primers designed following alignment of conserved region of characterised halophilic *Bacilli* proteases [16]. Secondly, while most of these studies claim novelty, most of the protease isolated actually share significant homology with already characterised proteases. For example the serine protease isolated by Biver and colleagues [11] is from a well characterised MEROPS [17] group of proteases; the S8 serine protease group or subtilase family and shares 69% homology with a subtilis-like proprotein convertase from *Desuflobacter postatei* (accession number WP_004070934). Similarly, the serine protease isolated by Pushpam and colleagues [9] shared 89% similarity with another S8 secreted peptidase A from *Shewanella* sp. (accession number HM370566). Perhaps it is not really feasible to call such proteases 'novel' as they only expand the current gene space and do not represent new sets within this space. Thus, the necessity for the re-evaluation of current functional metagenomics methods and the need to document the isolation and characterisation of proteases found using these techniques is apparent if we are to develop these methods and ultimately access the full repertoire of the microbial degradome.

Many studies utilise the methods of enriching for bacteria with certain metabolic capacity which is the trait or target under analysis before creating their metagenomic library. For example, Lussier *et al.*,[18] used a sequencing fed-batch reactor (SFBR) enriched for bacteria that were able to grow at alkaline pH at thermophilic temperatures

and were successful in isolating clones with lipase activity (however, it is also important to note that they intended to isolate proteases as well, but were unsuccessful). Similarly, Elend and colleagues [19] enriched for bacteria with lipolytic activities. They inoculated media, supplemented with mineral oil, with an industrial soil sample and were subsequently successful in isolating 2 novel esterase genes. Gabor *et al.*, [20] enriched for amidase genes by complimenting a leucine auxotrophy in the surrogate host. However, as alluded by Lussier *et al*., and Elend *et al.*[18, 19] this strategy of enrichment for particular species can lead to a loss in diversity and present much more of a challenge with some ecosystems where, for example, the majority of the community are strict anaerobes.

AIMS

In this research, an ecosystem was chosen with the hypothesis that it would be enriched for proteases; a dairy waste water treatment plant which clearly has high levels of casein.

The aims of this research were as follows:-

1. Sub-clone two putative protease sequences derived from a fosmid clone exhibiting protease activity derived from a previously constructed metagenomic library from DNA extracted from this waste treatment and assess the sub-clones for recapitulation of a protease phenotype utilising the robust media developed in chapter 3.0.

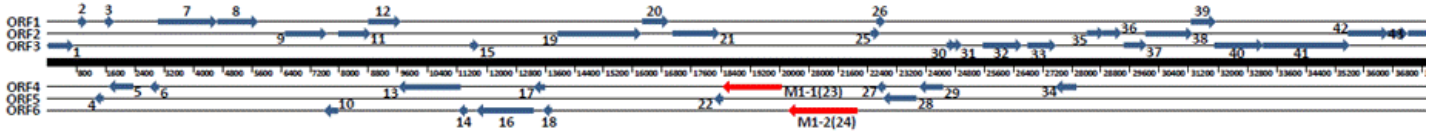
2. Conduct a comprehensive sequence analysis and comparison of the two derived putative protease sequences.

3. Conduct biochemical characterisation of the sub-clones.

4.2 RESULTS

4.2.1 Sequence analysis of M1-1 and M1-2 putative protease-encoding sequences

From the screening of approximately 28,000 fosmid clones, 1 positive metagenomic clone was isolated and sequenced. GeneMark software was used to deduce putative open reading frames within the protease positive 38 kb fosmid. Two open reading frames that potentially encoded the protease genes most likely responsible for the proteolytic phenotype observed were deduced by BLAST analysis and termed M1-1 (complement (18489..20057)) and M1-2 (complement (20297..22117))(Figure 4.1; Accession number for the fosmid is HF548541). M1-1 encoded a protein of 523 amino acids in length with a predicted molecular weight of 55.53 kDa. M1-2 encoded a protein composed of 606 amino acids with a predicted molecular weight of 63.79 kDa. Putative promoter sites were found in both M1-1 and M1-2 as were potential signal peptide regions.



CDS	Description	Total score	Query coverage	E value	Max ident	Positives
1	YP_004664826.1| non-ribosomal peptide synthase-like protein [Myxococcus fulvus HW-1]	75.1	94%	9e-13	31%	43%
2	YP_004057314.1| hemolysin a [Oceanithermus profundus DSM 14977]	32.7	40%	2.7	50%	81%
3	No significant similarity found.					
4	No significant similarity found.					
5	YP_006374191.1| GTPase domain-containing protein [Tistrella mobilis KA081020-065]	150	79%	3e-42	47%	63%
6	No significant similarity found.					
7	ZP_10292364.1| Cyclic peptide transporter [Pseudoalteromonas rubra ATCC 29570]	379	98%	9e-122	37%	60%
8	YP_005033682.1| protein of unknown function [Azospirillum brasilense Sp245]	220	82%	3e-65	39%	59%
9	EIJ45665.1| chorismate synthase [Herbaspirillum sp. GW103]	633	98%	0.0	83%	91%
10	EIT88516.1| hypothetical protein A200_02880 [Parascardovia denticolens IPLA 20019]	33.9	83%	3.8	31%	47%
11	ZP_04577216.1| Nucleoside:H+ symporter:Major facilitator superfamily MFS_1 [Oxalobacter formigenes HOxBLS]	246	99%	2e-76	50%	67%
12	CAL61391.1| putative Peptidase M48 [Herminiimonas arsenicoxydans]	332	96%	9e-111	58%	76%
13	ZP_10443015.1| hypothetical protein JPAM2_11455 [Janthinobacterium sp. PAMC 25724]	589	99%	0.0	53%	73%
14	YP_662217.1| ThiJ/PfpI [Pseudoalteromonas atlantica T6c]	30.4	82%	8.1	42%	67%
15	No significant similarity found.					
16	YP_004696293.1| hypothetical protein Nit79A3_3156 [Nitrosomonas sp. Is79A3]	499	90%	4e-170	54%	67%
17	YP_212851.1| histidyl-tRNA synthetase [Bacteroides fragilis NCTC 9343]	34.7	96%	1.2	27%	49%
18	EJL59593.1| site-specific recombinase XerD [Rhizobium sp. CF080]	31.2	54%	6.7	48%	76
19	YP_006214843.1| signal transduction histidine-protein kinase BaeS [Providencia stuartii MRSN 2154]	277	51%	2e-81	41%	59%
20	YP_004195100.1| winged helix family two component transcriptional regulator [Desulfobulbus propionicus DSM 2032]	250	99%	1e-80	56%	74%
21	YP_004754325.1| putative PAS/PAC sensor protein [Collimonas fungivorans Ter331]	378	65%	9e-127	69%	79%
22	ZP_05069698.1| conserved hypothetical protein [Candidatus Pelagibacter sp. ATCC7211]	30.8	82%	4.7	45%	58%
23	**NP_899726.1| aminopeptidase [Chromobacterium violaceum ATCC 12472]**	**446**	**74%**	**6e-150**	**55%**	**71%**
24	**YP_004752195.1| Neutral protease [Collimonas fungivorans Ter331]**	**746**	**100%**	**0.0**	**62%**	**76%**
25	ZP_06413597.1| Silent information regulator protein Sir2 [Frankia sp. EUN1f]	31.6	84%	4.6	38%	51%
26	No significant similarity found.					
27	YP_832751.1| nitrite reductase (NAD(P)H) large subunit [Arthrobacter sp. FB24]	33.9	63%	1.2	47%	67%
28	YP_437728.1| hypothetical protein HCH_06672 [Hahella chejuensis KCTC 2396]	139	64%	3e-36	39%	62%
29	NP_902234.1| hypothetical protein CV_2564 [Chromobacterium violaceum ATCC 12472]	209	84%	1e-65	61%	76%
30	YP_006055522.1| hypothetical protein SCATT_36290 [Streptomyces cattleya NRRL 8057 = DSM 46488]	31.6	75%	4.6	33%	52%
31	EJV42148.1| hypothetical protein IEK_05814 [Bacillus cereus BAG6O-1]	29.6	71%	5.6	50%	63%
32	YP_004754063.1| quinone oxidoreductase [Collimonas fungivorans Ter331]	452	98%	2e-156	65%	80%
33	YP_001100108.1| triosephosphate isomerase [Herminiimonas arsenicoxydans]	310	99%	1e-103	67%	78%
34	ZP_06708698.1| conserved hypothetical protein [Streptomyces sp. e14]	35.8	70%	2.0	29%	41%
35	YP_003775187.1| NADH dehydrogenase I subunit A [Herbaspirillum seropedicae SmR1]	218	85%	1e-70	88%	93%
36	YP_001353153.1| NADH dehydrogenase subunit B [Janthinobacterium sp. Marseille]	316	98%	2e-108	94%	99%
37	ZP_10722588.1| NADH dehydrogenase, subunit C [Herbaspirillum sp. CF444]	317	100%	2e-107	76%	84%
38	YP_001100103.1| NADH dehydrogenase subunit D [Herminiimonas arsenicoxydans]	819	100%	0.0	93%	97%
39	YP_001100102.1| NADH dehydrogenase subunit E [Herminiimonas arsenicoxydans]	295	75%	3e-99	84%	96%
40	ZP_10443842.1| NADH-ubiquinone oxidoreductase subunit F [Janthinobacterium sp. PAMC 25724]	794	100%	0.0	90%	94%
41	YP_001353158.1| NADH dehydrogenase I chain G [Janthinobacterium sp. Marseille]	1159	99%	0.0	73%	84%
42	YP_001100099.1| NADH dehydrogenase I subunit H [Herminiimonas arsenicoxydans]	560	100%	0.0	77%	86%
43	ZP_08275180.1| NADH-ubiquinone oxidoreductase chain I [Oxalobacteraceae bacterium IMCC9480]	301	100%	1e-102	93%	97%
44	ZP_10722595.1| NADH:ubiquinone oxidoreductase subunit 6 (chain J) [Herbaspirillum sp. CF444]	322	98%	6e-109	74%	87%
45	ZP_10722596.1| NADH:ubiquinone oxidoreductase subunit 11 or 4L (chain K) [Herbaspirillum sp. CF444]	187	100%	2e-59	93%	99%

1

← *From previous page*

Figure 4.2 Annotation of the fosmid insert from which the two putative proteases were identified and cloned. The table below the figure shows the nearest BLASTP matches to each ORF within the fosmid insert and the host species in square brackets.

M1-1 sequence comparison against GenBank showed M1-1 shared greatest sequence identity (55%) with a probable aminopeptidase from *Chromobacterium violaceum*, but also with leucyl aminopeptidases present in many members of the *Burkholderia* genus of *Proteobacteria* (Fig. 4.2). M1-1 contained a EE sequence at position 358-363 and the HXD (where X is any non-conserved amino acid) sequence at position 326-329 (Fig. 4.3a). Both sequences are highly conserved zinc ligand motifs associated with a family of metalloproteases as characterised by the MEROPS database as M28 or the aminopeptidase-like family of which an aminopeptidase of *Streptomyces griseus* is the prototype and includes both amino and carboxy-peptidases.

M1-2 also shared greatest similarity to a family of metalloproteases. The conserved regions found in M1-1 were absent in M1-2 (Fig. 4.3b). M1-2 was, however, found to contain a HEVSH sequence at position 353-366 and a GGINEAFSD sequence at positions 382-390 (Fig. 4.3b). HEXXH and GXXNEXXSD are highly conserved regions amongst zinc metalloproteases belonging to the M4 family the former being a zinc-binding motif, the latter, a third zinc ligand motif. M1-2, according to the BLAST database, was found to share greatest amino acid sequence similarity to a neutral protease from *Collimonas fungivorans* and to a class 4 metalloproteases from *Chromobacterium violaceum* (62% identity and 63% identity respectively) and also showed similarity to a number of proteases from the *Aeromonas* genus (Fig. 4.2) all of which belong to the M4/thermolysin family. It also showed a sequence similarity (56% identity) to a previously described metagenome-derived metalloprotease; MprA [21]. The majority of the ORFs clustered within the *Proteobacteria* and in the class β-*proteobacteria* (Figure 4.1).

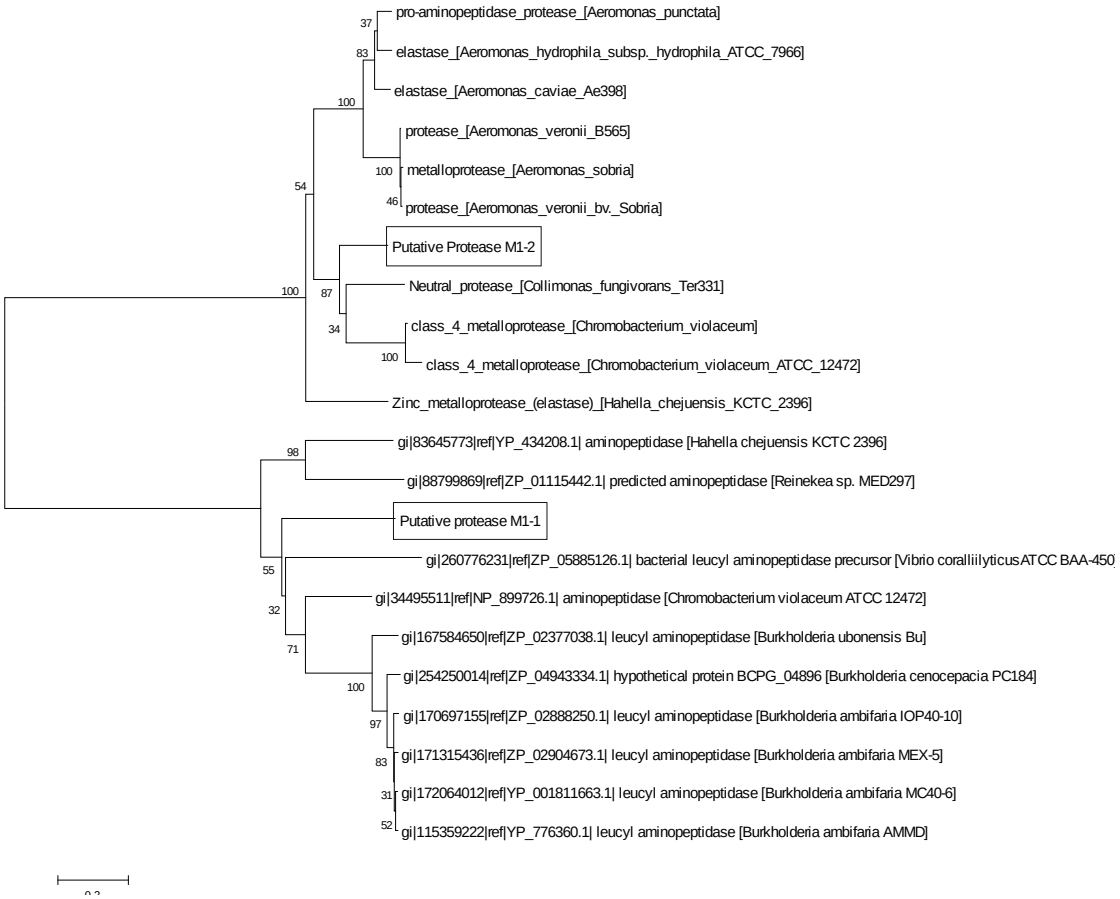


Figure 4.2 Phylogenetic tree of M1-1 and M1-2 sequences and their closest genetic matches as deduced from BLAST. This cluster analysis was based upon a neighbour-joining method as described in materials and methods. The tree includes the bacterial protease and the host organism name.

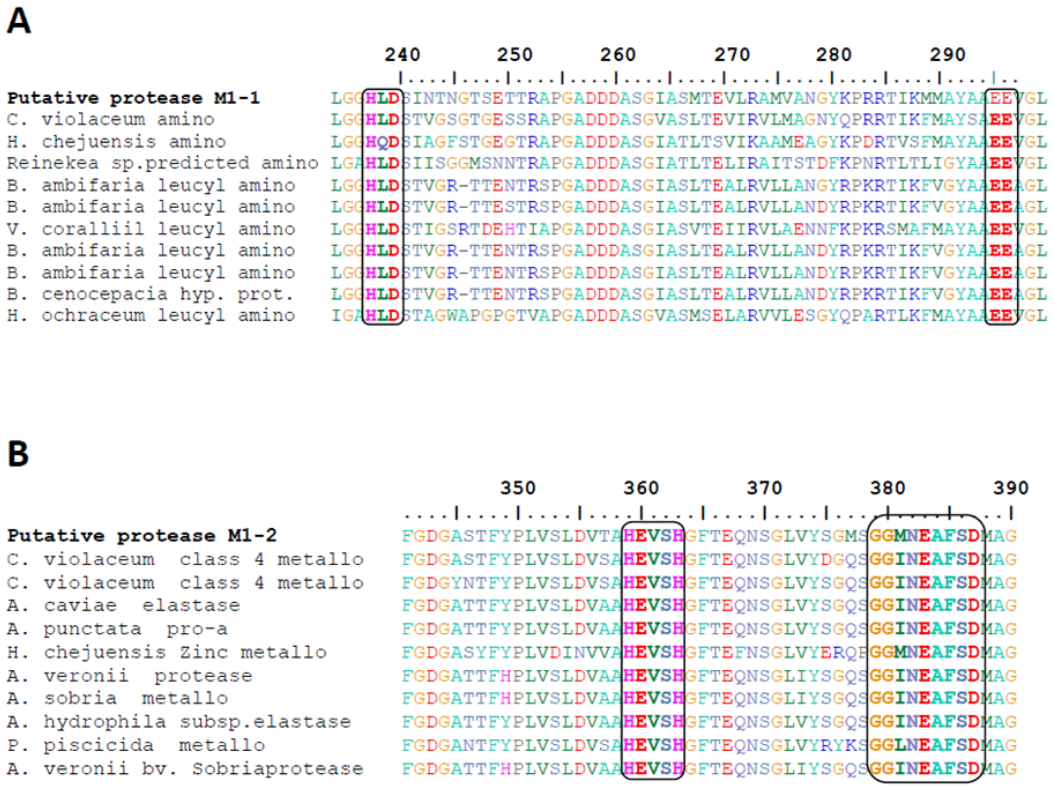


Figure 4.3. A Positions of the His-Leu-Asp (HLD) and Glu-Glu (EE) regions found in the M1-1 sequence and some of the closest genetic matches as deduced from BLAST, all of which are members of the M28 family of aminopeptidases in MEROPS database. **B:** Positions of the His-Glu-Val-Ser-His region (HEXXH) and Gly-Gly-Ile-Asn-Glu-Ala-Phe-Ser-Asp (GGXXNEXXSD) region found in M1-2 and some of the closest genetic matches deduced by BLAST all of which, including M1-2, belong to the M4 family of metalloproteases in the MEROPS database. Key to genus abbreviations; A – *Aeromonas*, B – *Burkholderia*, C – *Chromobacterium*, H – *Hahella*, and V – *Vibrio*.

Comparison with the PFAM database, MEROPS and CDD displayed very different enzyme structure and domain architecture (figure 4.4). The aminopeptidase M28 domain was at position 218-399 sharing homology to the same domain found in other members of the M28 family as also contained the regions for predicted active sites. M1-1 also had a pre-peptidase domain at the C terminus, a sequence that is not present in the active protease as it gets cleaved upon secretion. M1-2 was found to have a much more complex modular structure consisting of a thermolysin protein precursor at the N-terminus which is hypothesised to be involved in inhibiting premature activation of the enzyme [22]. The thermolysin region is highly conserved amongst the M1-2 BLAST sequences mentioned previously (figure 4.2). The thermolysin region was followed by a PepSY domain, a region believed to have inhibitory and chaperone properties [23, 24] and the catalytic domains containing the active site and zinc binding sites. Similarly to M1-1, M1-2 had a pre-peptidase C-terminal domain that was not part of the active protease. In addition using PFAM, 3 domains were identified for M1-2, which correspond to the protein fragments shown in the induced lane in Figure 4.5.

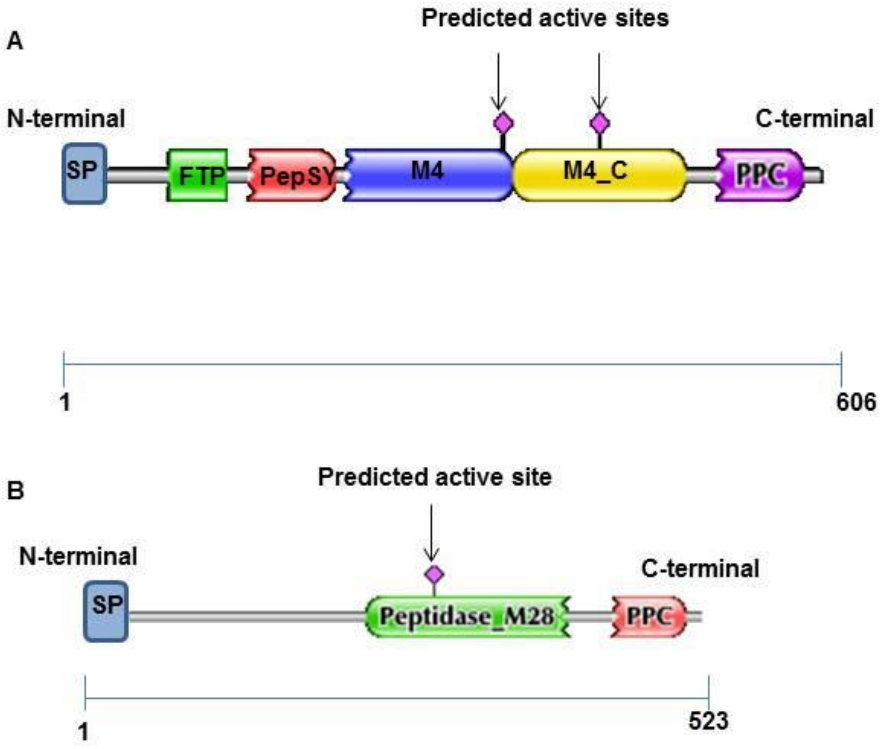


Figure 4.4 The domain architecture of; A M1-2 and B putative protease M1-1. SP: signal peptide, FTP: fungalysin/thermolysin propeptide, PepSY; peptidase domain (M4) and YpeB of *B. subtilis*, PPC; pre-peptidase C-terminal domain.

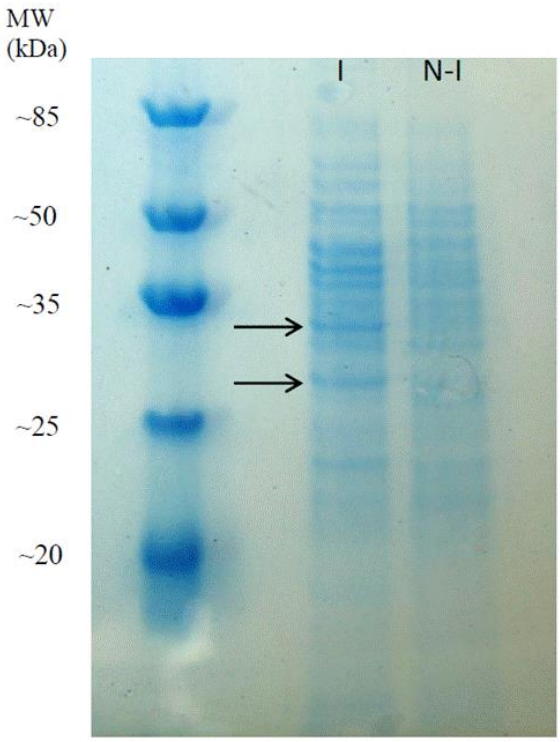


Figure 4.5 SDS-PAGE of the cell free extracts of *E. coli* Xl1-Blue cells harbouring the pGEM ®T-Easy vector with M1-2 sequence after IPTG induction and non-induced samples. SDS-PAGE was performed using the Laemmli method with a 12% w/v polyacrylamide gel. Gels were run with protein standard (Prestained Protein Molecular Weight Marker 20 to 120 kDa) and stained with InstantBLue™ Coomassie based protein staining solution (Fermentas, UK). Black arrows indicate bands present in the induced sample and therefore representing the cleaved region of M1-2 (B ~30kDa) and the active protease region of M1-2 (A ~32kDa)

4.2.2 Activity based screening of sub clones

Upon sub-cloning, transformants were screened on LF-SMA to assess the recapitulation of the protease phenotype on a robust agar. M1-2 was able to recapitulate this phenotype, however, M1-1 did not. The SDS-PAGE on cell-free extract on M1-2 exhibited two bands present in the induced and not present in the non-induced samples with apparent molecular masses of ≈32 kDa and ≈21kDa (Figure 4.5).

4.2.3 Characterisation of M1-2 protease

The optimal temperature for M1-2 activity was 42°C after incubation for 4 hours (Fig. 4.6a). Activity was below 30% of maximum at temperatures of 4, 10, 20 and 60°C and lower than 50% activity was observed at 50°C. The optimal pH for M1-2 activity was 8.0 (Fig. 4.6b).

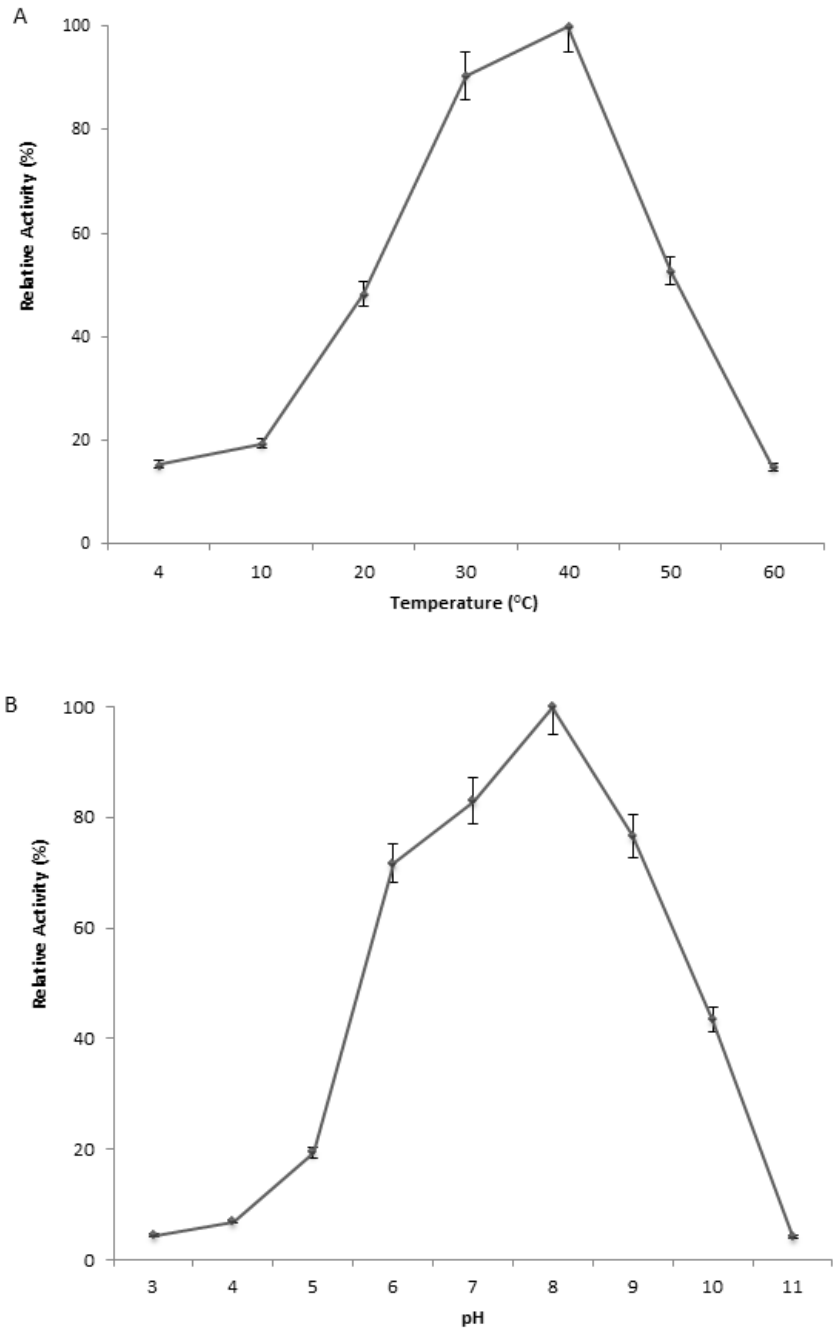


Figure 4.6 A: The Effect of temperature on protease activity as measured by the release of acid-soluble substance from azocasein (5 mg/ml) in HCl buffered Tris to a pH of 8.0 incubated at 4°C, 10°C, 20°C, 30°C, 40°C, 50°C and 60°C for 4 hours. Results are shown as the mean value of the results that were in triplicate and are shown as a relative

← *continued from previous page*

percentage of protease activity. **B**: The Effect of pH on protease activity as measured by the release of acid-soluble substance from azocasein (5 mg/ml) at varying pHs using sodium acetate for pH 3.0-5.0, Tris-HCl for pH 6.0-8.0 and glycine-NaOH for pH 9.0-11.0 incubated at 37°C for 4 hours. Results are shown as the mean value of the results that were in triplicate and are shown as a relative percentage of protease activity.

The effects of the metal ions and reagents on M1-2 activity described in experimental procedures can be seen in Table 4.1. Activity was slightly enhanced upon the addition of Mg^{2+}, Ni^{2+} and Co^{2+}, but inhibited by the addition of Zn^{2+}. Both 1 mM and 10 mM concentrations of the metal ion chelator EDTA reduced activity to below 30% activity and 1 mM SDS reduced activity to 35%. When 50 mM Zn^{2+} was added to 10 mM EDTA activity began to increase to >30%, but at lower concentrations (1 mM EDTA and 10 mM $ZnCl_2$) the protease was further inhibited.

Table 4.1: The effects of metal ions and chemical reagents on protease activity as described in Material and Methods. Effect was measured using the release of acid-soluble substance from azocasein (5 mg/ml) in HCl buffered Tris to a pH of 8.0 incubated at 37°C for 4 hours. Results are shown as the mean value of the results that were in triplicate and are shown as a relative percentage of protease activity.

Metal ion or reagent	Concentration (mM)	% Relative Activity
None	**None**	**100**
EDTA	**1**	**26**
EDTA	**10**	**24**
EDTA / ZnCl2	**1 mM EDTA /10 mM ZnCl2**	**12**
EDTA / ZnCl2	**10 mM EDTA / 50 mM ZnCl2**	**32**
CaCl2	**1**	**97**
ZnCl2	**1**	**19**
MgCl2	**1**	**104**
CoCl2	**1**	**102**
NiSO4	**1**	**105**
SDS	**1**	**35**

4.3 DISCUSSION

Functional metagenomics provides the potential to access some of the currently vastly uncharacterised or unknown microbial resources from any environment by-passing the need for laboratory culture techniques. Its potential has been demonstrated so far with the identification of novel proteins, antibiotics and antibiotic resistant genes [25, 26], however, the frequency of successful functional metagenomic screens is still very low as there are currently a number of limitations particularly when screening for proteolytic clones as was the aim of this research. These limitations include the possibility that a foreign protease gene may exert a toxic effect on surrogate host proteins possibly resulting in host cell death thus hindering protease recovery and also as the chosen surrogate host may not be capable of transcribing and/or translating this foreign protein leading to no active product. A further current limitation of functional metagenomics is the insufficient screening means [8]. Any steps towards mitigating these issues can only be seen as a positive for the future of metagenomics.

From our study, Valio™ lactose-free milk agar was found to be an effective and robust agar for correctly identifying proteases by way of distinct zones of clearing around a bacterial colony. From the screens undertaken in this study no false positives arose with the use of this type of agar. We would suggest that this agar would be much more appropriate for future screening of metagenomic libraries for protease activity. The robustness of this agar has and will allow for more efficient characterisation of enzyme activity which can lead to identification of clinically or industrially relevant proteases in the future. Here, a metagenomic library was created from the metagenome of a milk waste treatment plant in Cork, Ireland in a large fosmid vector with *E. coli* as the surrogate host. Following screening of the subsequent metagenomic clones on normal SMA one fosmid was found to confer a proteolytic phenotype. Sequence analysis of this 38kb fosmid revealed two putative protease sequences that may have been responsible for this phenotype. The sequences were analysed to find signal peptides, potential promoter sites, domains and motifs (figure 4.4) and were also aligned with their closest genetic matches as determined by Protein BLAST searches to deduce phylogeny and to determine to which protease family they belonged. The sequences were subcloned into another strain of *E.coli* which was found to be capable of synthesising and secreting one of the two foreign protease sequences found, exemplifying both the usefulness, but also

limitations of *E.coli* as the standard host for cloning. LF-SMA was used to avoid the problem described by Jones and co-workers [8] and was successful in identifying M1-2 as recapitulating a proteolytic phenotype, allowing subsequent characterisation and identification. An alternative vector; pET-42a had to be implemented for M1-1 following no recapitulation of a protease phenotype in pGEM®T-easy. This vector is useful for the quick cloning and high level expression of protein in a host bacterial cell. However, this method still proved fruitless and no phenotype was observed though it was still deemed necessary to analyse the sequence and phylogeny to either disrepute this sequence as a protease, or to provide answers for the lack of expression to aid the progress of metagenomics.

Sequence analysis revealed both sequences to be extracellular zinc-dependent metalloproteases which are a diverse group of proteases. Phylogenetic analysis using a neighbour-joining method concluded M1-1 and M1-2 to be phylogenetically distinct demonstrating the varying evolutionary lineages of metalloproteases and was in accordance with the family separation of the two proteases as exhibited by the MEROPS database and the dissimilar modular structures suggesting the sequences did not come from the same organism.

M1-2 was shown to belong to the highly heterogeneous M4 family of proteases. The closest genetic matches of M1-2 include metalloproteases from *Chromobacterium violaceum, Collimonas fungivorans* and *Aeromonas* species. M1-2 was also found to share 56% sequence identity to a previously metagenome-derived metalloprotease found in soil [13] Since the closest identity to M1-2 include proteases from soil inhabitants, water inhabitants it can be suggested that this M1-2 may provide the genetic capability to utilise a diverse array of substrates from different ecological niches. It is well established that waste treatment plants provide an intermediary for horizontal gene transfer [27],[28] so such genetic material may have easily passed from one microorganism to another. Since the closest genetic matches to M1-2 are also ubiquitous in soil and aquatic environments there is good chance of horizontal gene transfer of genetic material in an environment such as a waste treatment plant. It is also important to note that of the closest genetic matches for M1-2 all proteases came from members of the *Proteobacteria* which may be an indicator as to why *E. coli* was able to express M1-2 successfully. This potential inability of *E. coli* to express M1-1 but it's ability to

express M1-2 and the information we have learnt from sequence analysis of M1-2 has useful implications for the host selection for future metagenomic screens, since phylogeny can be a good indicator of genetic capacity and whether or not it would be a suitable surrogate host for the expression of DNA from the environment under scrutiny.

M4 family characterisation is based on the protease thermolysin from *Bacillus thermoproteolytcius*, thus members of this family are generally metalloendopeptidases that exhibit a degree of thermotolerance. Characterisation of M1-2 temperature optima is commensurate with this thermostable characteristic that was comparable to the tolerance shown by members of the M4 family [29] M4 require Zn^{2+} ions for catalytic activity and Ca^{2+} ions for stability. M1-2 activity was unsurprisingly inhibited when the ion chelator EDTA was added and activity was slightly re-instated upon the addition of $ZnCl_2$ and $CaCl_2$ (Table 2). Although, surprisingly activity was decreased to 19% (Table 2) in the presence of an excess of zinc ions though previous research has shown that an excess of the corresponding ion needed for activity (in this instance, zinc) activity was reduced due to over-binding of the ion to the active site consequently inhibiting activity [30].

As they do require Zn^{2+} ions, though not in excess, for activity, all members of the M4 family contain the amino acid zinc-binding motif: HEXXH which was present in the sequence of M1-2 and its closet genetic matches as deduced by a Protein BLAST search all of which were metalloproteases belonging to the M4 family. Although they are a highly heterogeneous group, the catalytic site amongst M4 proteases is actually highly homogeneous. M1-2 putative catalytic site was shown to share (98%) identity with other members of M4 thus we can confirm that the M1-2 sequence is a zinc-dependent metalloprotease that is the most credible candidate responsible for the proteolytic phenotype observed from the first screen of the metagenomic library and also that at least the catalytic domain must be ubiquitous throughout many environmental bacteria.

M1-1 may not have been responsible for the proteolytic phenotype initially observed by the metagenomic fosmid; it still cannot be disregarded as a protease as many of the related proteases were leucyl aminopeptidases from *Burkholderia* (Fig. 3) suggesting *E.coli* may not have been capable of synthesising this gene. Thus, regardless of whether or not the M1-1 sequence confers a protease, this research has highlighted that in order to access the full repertoire of extracellular microbial protease function, it is may not be

sufficient to rely on current standard screening techniques nor *E. coli* as the standard host and since a similar sequence to M1-2 has been previously isolated [13] the functional metagenomic techniques currently used may favour the expression of certain types of proteases over others indicating that it would be advantageous to develop metagenomic methods that allow for an expansion of host range for more competent expression of novel genes.

Recent culture-independent analysis by deep-sequencing approaches has enabled us to recognise the genetic diversity and potential of the gut microbiota. However, identifying functional characteristics from metagenomes could be more helpful in deducing microbial-host interactions and the impact they have with regards to pathogenesis and disease aetiology [31]. In this instance, we were interested in the role of proteases for reasons described previously. The ultimate aim of this research is to understand proteolytic functions within the human gut microbial metagenome as within the human gut, microbial proteases have been largely untapped as resources, and have been implicated with the pathology of inflammatory bowel disorders, diarrhoeal disease and colon cancer. There are many other proteolytically significant bacteria residing in the human gut of which the effects of proteolysis on the host are largely unknown. It is also likely that secondary metabolites as a result of protein degradation in the human gut by the gut microbiota leads to a build up of toxic compounds such as phenolic compounds, ammonia and amines [32]. Hence we have the motivation for attempting to further uncover the degradome of the gut microbiota which will be achieved by culture-dependent analysis of proteolysis in the gut microbiota to deduce diversity and taxonomy, and improving and developing the standard techniques that are currently implemented, and proving fairly unproductive for protease screening. Our novel techniques are based on the development and utilisation of a Gram positive host; *Bacillus subtilis s*train WB800N which is protease deficient but chosen as a suitable surrogate host due to its GRAS (generally regarded as safe) status and the ability of the wild type to produce and secrete numerous proteases. A Gram negative host (*E. coli*) will be used for maintenance as the pNZ44 vector currently being used for library construction (Chapter 5.0) has both Gram positive and Gram negative replication origins Libraries can be initially screened on our LF-SMA developed in this research (Chapter 3.0).

4.4 MATERIAL AND METHODS

4.4.1 Bacterial strains, plasmids and growth conditions

The strains and plasmids used throughout this study along with conditions for growth/use are listed in Chapter 2.0, table 2.1. Plasmid and fosmid preparations on *E. coli* were carried out using the Qiaprep spin mini prep kit (QIAGEN LTD. West Sussex, UK) according to the manufacturer's protocol on cells from an overnight culture. A fosmid library (28032 clones (pCC1Fos), in *E. coli* EPI300) was constructed as described previously [8, 33] using high molecular weight DNA isolated from a settling tank's activated sludge. This library was screened on skimmed milked agar, containing chloramphenicol (12·5 μg ml^{-1}) as described in [14]. Since false positives are a possibility of this assay [8] positive clones were further tested to ensure they did not produce acetic acid and were able to degraded casein.

The fosmid insert(s) from any positive clones was sequenced by MWG (Ebersberg, Germany) and GeneMark [34] used to predict the putative open reading frames, which were annotated using the BLAST suite of tools [35]. Putative protease were compared to the MEROPS peptidase database [36] and signal peptides identified using the SignalP 3.0 Server [37]. Promoter prediction was determined using the Promoter 2.0 Prediction software [38]. Potential structures and domains were compared to sequences in the Conserved Domain Database (CDD) [39] and Pfam database [40]. Molecular evolutionary analysis of the two putative protease sequences and their ten closest protease matches following the BLAST analysis was inferred using Neighbor-Joining algorithms [41] in MEGA version 5.0 [42]. Taxonomic analysis of the ORFs was determined using the web-server WebMGA [43].

PCR reactions to amplify protease sequences from the fosmid were performed on a C1000™ Thermal cycler. Primers were designed to amplify the open reading frames. Reactions were routinely 25μl prepared with 0.2 μM primers (M1-1 forward;5'-GGACAGTTCGCGCAAGCCCT-3', M1-1 reverse; 5'-TGCCATGCTTCATGCTAGAAATGC-3', M1-2 forward; 5'-AGGATATAGAAAGCATCCCG-3' and M1-2 reverse; 5'-GCTGGGGCGGTTCTGCGTAA-3'), 1U of NEB *Taq* DNA Polymerase (New England Biolabs® Inc. Ipswich), 1X standard reaction buffer (10 mM Tris-HCl, 50 mM KCl, 1.5

mM $MgCl_2$, pH 8.3 at 25°C), 1X Bovine Serum Albumin (NEB), 0.2 µM dNTPS, sterile H_2O and template DNA (≈10 ng fosmid DNA). PCR conditions were as follows: Heating at 94°C for 5 mins followed by 30 cycles of heating 94°C for 30s, 60°C for 45s and 72°Cfor 2 mins and a final step of 72°C for 7 mins. PCR products were analysed on a 0.8% (w/v) gel with ethidium bromide (1 µg/ml). PCR products were purified using the Qiagen PCR purification kit according to their protocol. For products that were to be cloned into pET-42a, the following primers were used to include *Nde*I and *Xho*I sites and PCR conditions were as described as above M1-1 forward;5'-GGCCTCGAGGGACAGTTCGCGCAAGCCCT-3', M1-1 reverse; 5'-TGCCATGCTTCATGCTAGAAATGC-3', M1-2 forward; 5'-AGGATATAGAAAGCATCCCGGTATACGGC-3'

4.4.2 Sub-cloning and activity based screening

Following amplification by PCR of the putative protease open reading frames and flanking regions to include potential promoter sites the purified inserts were treated with T4 DNA polymerase (Fermentas, UK) according to the manufacturer's protocol in order to fill in any 5' or 3' overhangs. The insert was ligated into the pGEM®T-easy (Promega, Southampton, UK) at a 3:1 insert to vector ratio using T4 DNA ligase (Promega). The subsequent ligation mixture (2µl) was used to transform *E. coli* XL1-Blue supercompetent cells (Agilent, Wokingham, UK) according to the manufacturer's protocol. Following identification of transformants positive for the insert by blue-white screening on LB plates supplemented with 0.5 mM IPTG and 100 µg/ml ampicillin, white colonies were re-screened on a lactose free skimmed milk agar (LF-SMA) which was prepared as previously described [44] supplemented with the appropriate antibiotic.

4.4.3 Preparation of cell free extract and enzyme characterisation

Cells were grown overnight in LB media supplemented with the 100 µg/ml ampicillin. 1% (v/v) of this starting culture was used to inoculate 50 ml of fresh media and grown to an optical density OD_{600} = 0.5. The culture was divided into a non-induced control and IPTG induced culture in which IPTG was added to a final concentration of 0.5 mM and both were left to grow for a further 3 hours. Cells were harvested by centrifugation at 1500 x g for 10 mins. The pellet was re-suspended in 2 ml of phosphate buffered saline (PBS) and each sample was bead beaten for 30 secs, 3 times with 0.5g of 0.1mm glass

beads. Samples were centrifuged at 20,000 x g for 10 mins and the resulting supernatant was taken as the cell free extract. Protein was quantified by the method of Bradford with bovine serum albumin as the standard [45]. Sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) was performed according to Laemmli 1970 on a 12% w/v gel [46].

Protease activity was determined by measuring the release of acid-soluble substance from azocasein (Sigma-aldrich, Poole, UK) as described in the general materials and methods section (Chapter 2.0 section 2.9.1)

In order to determine the optimum temperature for protease activity, the previously described assay was carried out, but incubated at the following temperatures; 4°C, 10°C, 20°C, 30°C, 40°C, 50°C and 60°C for a period of 4 hours. The reactions were terminated by addition of 10% (w/v) TCA and protein precipitated as described above. Similarly, to deduce pH optima, the pH of the azocasein solution was varied from 3 to 11 using a sodium acetate (pH range 3-5), Tris-HCL (pH range 6-8) or glycine sodium hydroxide (pH range 9-11) buffer system.
To investigate the effect of specific metal ions and chemical reagents, 100 µl of protease sample was incubated for 3 h with 100 µl azocasein in the presence of varying concentrations of ethylenediaminetetraacetic acid (EDTA), sodium dodecyl sulphate (SDS) $CaCl_2$, $CoCl_2$, $NiSO_4$ and $ZnCl_2$.

4.5 REFERENCES

1. Morris, L.S., Evans, J., and Marchesi, J.R. (2012). A robust plate assay for detection of extracellular microbial protease activity in metagenomic screens and pure cultures. Journal of Microbiological Methods *91*, 144-146.
2. Rao, M.B., Tanksale, A.M., Ghatge, M.S., and Deshpande, V.V. (1998). Molecular and Biotechnological Aspects of Microbial Proteases. Microbiology and Molecular Biology Reviews. *62*, 597-635.
3. Ten, L.N., Im, W.T., Kim, M.K., and Lee, S.T. (2005). A plate assay for simultaneous screening of polysaccharide- and protein-degrading micro-organisms. Letters in Applied Microbiology *40*, 92-98.
4. Maeda, H., and Molla, A. (1989). Pathogenic potentials of bacterial proteases. Clinica Chimica Acta *185*, 357-367.
5. Lantz, M.S. (1997). Are bacterial proteases important virulence factors? Journal of Periodontal Research *32*, 126-132.
6. Shen, H.-B., and Chou, K.-C. (2009). Identification of proteases and their types. Analytical Biochemistry *385*, 153-160.
7. Jones, B.V., Begley, M.i., Hill, C., Gahan, C.G.M., and Marchesi, J.R. (2008). Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. Proceedings of the National Academy of Sciences *105*, 13580-13585.
8. Jones, B.V., Sun, F., and Marchesi, J.R. (2007). Using skimmed milk agar to functionally screen a gut metagenomic library for proteases may lead to false positives. Letters in Applied Microbiology *45*, 418-420.
9. Pushpam, L., Rajesh, T., and Gunasekaran, P. (2011). Identification and characterization of alkaline serine protease from goat skin surface metagenome. AMB Express *1*, 1-10.
10. El Said, N.H. (2012). Identification of potential subtilisin-like serine proteases from the microbial community of the Red Sea Atlantis-II brine pool using a metagenomic approach. In Department of Biology, Volume Biotechnology. (American University of Cairo).
11. Sophie Biver, D.P.a.M.V. (2013). Characterization of a new oxidant-stable serine protease isolated by functional metagenomics. SpringerPlus *2*.
12. Neveu, J., Regeard, C., and DuBow, M. (2011). Isolation and characterization of two serine proteases from metagenomic libraries of the Gobi and Death Valley deserts. Applied Microbiology and Biotechnology *91*, 635-644.
13. Waschkowitz, T., Rockstroh, S., and Daniel, R. (2009). Isolation and Characterization of Metalloproteases with a Novel Domain Structure by Construction and Screening of Metagenomic Libraries in Applied and Environmental Microbiology. *75*, 2506-2516.
14. Berlemont, R., Delsaute, M., Pipers, D., D'Amico, S., Feller, G., Galleni, M., and Power, P. (2009). Insights into bacterial cellulose biosynthesis by functional metagenomics on Antarctic soil samples. ISME J *3*, 1070-1081.
15. Zhang, Y., Zhao, J., and Zeng, R. (2011). Expression and characterization of a novel mesophilic protease from metagenomic library derived from Antarctic coastal sediment. Extremophiles *15*, 23-29.
16. Purohit, M.K., and Singh, S.P. (2013). A metagenomic alkaline protease from saline habitat: Cloning, over-expression and functional attributes. International Journal of Biological Macromolecules *53*, 138-143.

17. Rawlings, N.D., Barrett, A.J., and Bateman, A. (2104). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Research *40*, D343-D350.
18. Lussier, F.-X., Chambenoit, O., Côté, A., Hupé, J.-F., Denis, F., Juteau, P., Beaudet, R., and Shareck, F. (2010). Construction and functional screening of a metagenomic library using a T7 RNA polymerase-based expression cosmid vector. Journal of Industrial Microbiology and Biotechnology, 1-8.
19. Elend, C., Schmeisser, C., Leggewie, C., Babiak, P., Carballeira, J.D., Steele, H.L., Reymond, J.-L., Jaeger, K.-E., and Streit, W.R. (2006). Isolation and Biochemical Characterization of Two Novel Metagenome-Derived Esterases. Applied and Environmental Microbiology *72*, 3637-3645.
20. Gabor, E.M., De Vries, E.J., and Janssen, D.B. (2004). Construction, characterization, and use of small-insert gene banks of DNA isolated from soil and enrichment cultures for the recovery of novel amidases. Environmental Microbiology *6*, 948-958.
21. Waschkowitz, T., Rockstroh, S., and Daniel, R. (2009). Isolation and Characterization of Metalloproteases with a Novel Domain Structure by Construction and Screening of Metagenomic Libraries. Applied and Environmental Microbiology *75*, 2506-2516.
22. Markaryan, A., Lee, J.D., Sirakova, T.D., and Kolattukudy, P.E. (1996). Specific inhibition of mature fungal serine proteinases and metalloproteinases by their propeptides. Journal of Bacteriology *178*, 2211-2215.
23. Yeats, C., Rawlings, N.D., and Bateman, A. (2004). The PepSY domain: a regulator of peptidase activity in the microbial environment? Trends in Biochemical Sciences *29*, 169-172.
24. Marie-Claire, C., Roques, B.P., and Beaumont, A. (1998). Intramolecular processing of prothermolysin. Journal of Biological Chemistry *273*, 5697-5701.
25. Uchiyama, T., and Miyazaki, K. (2009). Functional metagenomics for enzyme discovery: challenges to efficient screening. Current Opinion in Biotechnology *20*, 616-622.
26. Allen, H.K., Donato, J., Wang, H.H., Cloud-Hansen, K.A., Davies, J., and Handelsman, J.(2010). Call of the wild: antibiotic resistance genes in natural environments. Nature Reviews Microbiology *8*, 251-259.
27. Barberio, C., Pagliai, L., Cavalieri, D., and Fani, R. (2001). Biodiversity and horizontal gene transfer in culturable bacteria isolated from activated sludge enriched in nonylphenol ethoxylates. Research in Microbiology *152*, 105-112.
28. Dröge, M., Pühler, A., and Selbitschka, W. (1999). Horizontal gene transfer among bacteria in terrestrial and aquatic habitats as assessed by microcosm and field studies. Biology and Fertility of Soils *29*, 221-245.
29. de Kreij, A., Venema, G., and van den Burg, B. (2000). Substrate Specificity in the Highly Heterogeneous M4 Peptidase Family Is Determined by a Small Subset of Amino Acids. Journal of Biological Chemistry *275*, 31115-31120.
30. Lee, S.H., Taguchi, H., Yoshimura, E., Minagawa, E., Kaminogawa, S., Ohta, T., and Matsuzawa, H. (1994). Carboxypeptidase Taq, a thermostable zinc enzyme, from *Thermus aquaticus* YT-1: molecular cloning, sequencing, and expression of the encoding gene in *Escherichia coli*. Biosciences Biotechnology and Biochemistry. 1994 Aug;58(8):1490-5.
31. Flores, R., Shi, J., Gail, M.H., Gajer, P., Ravel, J., and Goedert, J.J. (2012). Association of Fecal Microbial Diversity and Taxonomy with Selected Enzymatic Functions. PLoS ONE *7*, e39745.

32. Hughes, R., Magee, E.A., and Bingham, S. (2000). Protein degradation in the large intestine: relevance to colorectal cancer. Current Issues in Intestinal Microbiology *1*, 51-58.
33. Jones, B.V., Begley, M., Hill, C., Gahan, C.G.M., and Marchesi, J.R. (2008). Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. Proceedings of the National Academy of Sciences of the United States of America *105*, 13580-13585.
34. Lukashin, A.V., and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. Nucleic Acids Research *26*, 1107-1115.
35. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of Molecular Biology *215*, 403-410.
36. Rawlings, N.D., Barrett, A.J., and Bateman, A. (2012). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Research *40*, D343-D350.
37. Emanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. Nature Protocols *2*, 953-971.
38. Knudsen, S. (1999). Promoter2.0: for the recognition of PolII promoter sequences. Bioinformatics *15*, 356-361.
39. Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., et al. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Research *39*, D225-D229.
40. Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., et al. (2012). The Pfam protein families database. Nucleic Acids Research *40*, D290-D301.
41. Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution *4*, 406-425.
42. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution.
43. Wu, S., Zhu, Z., Fu, L., Niu, B., and Li, W. (2011). WebMGA: a customizable web server for fast metagenomic sequence analysis. BMC Genomics *12*, 444.
44. Morris, L.S., Evans, J., and Marchesi, J.R. (2012). A robust plate assay for detection of extracellular microbial protease activity in metagenomic screens and pure cultures. Journal of Microbiological Methods *91*, 144-146.
45. Bradford, M.M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Analytical Biochemistry *72*, 248-254.
46. Laemmli, U.K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature *227*, 680-685.
47. Morris, L.S.and Marchesi, J.R. (2014).Current functional metagenomic approaches only expand the existing protease sequence space, but does not presently add any novelty to it. Current Microbiology *284*

5.0 DEVELOPMENT OF FUNCTIONAL METAGENOMIC TOOLS AND SCREENING THE GUT MICROBIOME FOR PROTEASE ACTIVITY

5.1 INTRODUCTION

We now understand that while bacterial culture methods are extremely useful in discerning phenotypic characteristics of microorganisms, current techniques are not substantial to support the growth of the vast majority of microorganisms from a given environment. Our attempts to assess the microbial diversity via culture methods leads to a bias view of the bacteria that are actually present; those that grow are usually those that are the most successful at growing on nutrient rich, synthetic media, these successful growers or 'weeds' as they are commonly referred [1] are rarely the numerically dominant species in that particular environment [2]. This culture bias has become known as 'The great plate count anomaly'[2].

Polymicrobial infections such as that found in cystic fibrosis (CF) sufferers are a key example to show how culture-dependent diagnostics can overlook important pathogens. In this example, the most commonly isolated bacteria through routine aerobic culture are *Staphylococcus.aureus*, *Haemophilus influenza* , *Burkholderia cepacia* and *Pseudomonas aeruginosa* and were therefore thought to be the key players in infection [3]. However, recent molecular based studies have shown that the CF lung has a much greater species richness and many inhabitants are anaerobic such as *Prevotella* spp. and *Streptococcus* spp. and therefore would have been overlooked using standard aerobic culture conditions [4]. Some of these anaerobes may be of clinical significance [5] and therefore may have implications for antimicrobial therapy of pulmonary infection in CF patients. This example highlights how significant the issue of culture bias can be.

The human gut microbiota, a microbial habitat discussed at length previously (Introduction, Chapter 1.0) is another notable example of where culture bias can influence how we view the microbial community. In the early studies of the human gut microbiota researchers such as Finegold and colleagues [6] found only 113 difference bacterial species in the gut compared to the excess of 1000 species that recent culture-

independent sequencing technologies have now revealed there to actually be [7]. Culture-independent studies are therefore the favoured approach these days for studying microbial diversity and density of a given environment.

The emergence of Metagenomics, a term first coined by Handelsman and colleagues [8] provided a great opportunity to overcome the culture bias by directly studying the entire genome, or metagenome of the bacteria taken from a particulate environment be it soil, deep sea sediment, coral reef, hot springs or the human gut to name but a few. The resulting transformants can be analysed by one of two ways: sequencing analysis or functional analysis.

Functional metagenomics (FM) circumvents some of the limitations associated with shotgun metagenomic sequencing by providing a form of phenotypic identification; metagenomic clones are screened for a specific function identified by a measurable phenotype. FM therefore relies on the heterologous gene expression of the random metagenomic insert, and subsequent high-throughput screening for the desired functional characteristic. With this methodology, no prior genetic knowledge is required and so it is currently the only method that enables the identification of completely novel genes [9], making it an invaluable tool particularly when applied to complex and limitedly characterised environments. It has been shown to have the capabilities of identifying many novel genes, including those of medical significance for example Courtois et al. screened a metagenomic library generated from soil microbial DNA and unearthed novel polyketide antibiotic genes [10] research such as this helps prove FM to be a promising approach for drug discovery, which is of particular significance in an age where the rate of antibiotic resistance amongst microorganisms is faster than the discovery of new drugs to combat them. Conversely, Allen et al. implemented FM to Alaskan soil samples and discovered the first example of a bi-functional β-lactamase gene [11] providing insight into the extent of antibiotic resistance without the selection pressures of humans. Thus, understanding functions of microorganisms aids us in understanding how they have adapted to life in a particular environment and also allows the stimulation of new hypotheses [12]. Clearly, this process shows potential for not only furthering our knowledge of the microbial world, but also for effective drug discovery, and unravelling resistance mechanisms and mechanisms of disease.

However, despite the many advantages of FM, when compared to metagenomic sequencing, the process is far more intensive with the frequency of a positive hit remaining low; for a successful metagenomic screen a gene has to be successfully ligated into a vector, transformed into a surrogate host where it must be transcribed and translated into the fully functioning protein which must be secreted from the cell and its function demonstrated [9]. In many cases, clone libraries of tens of thousands of clones must be screened to get enough coverage (enough clones to represent the amoutnof DNA in a particular sample) in order to be able to identify clones with a positive phenotype. Perhaps the skewed ratio between effort put in and data generated is the reason for the current relative shortage of FM publications when compared to those related to sequencing metagenomics(SM) since here, an expansion of reads of genetic information can be generated, cost and time effective using high-throughput sequencing.

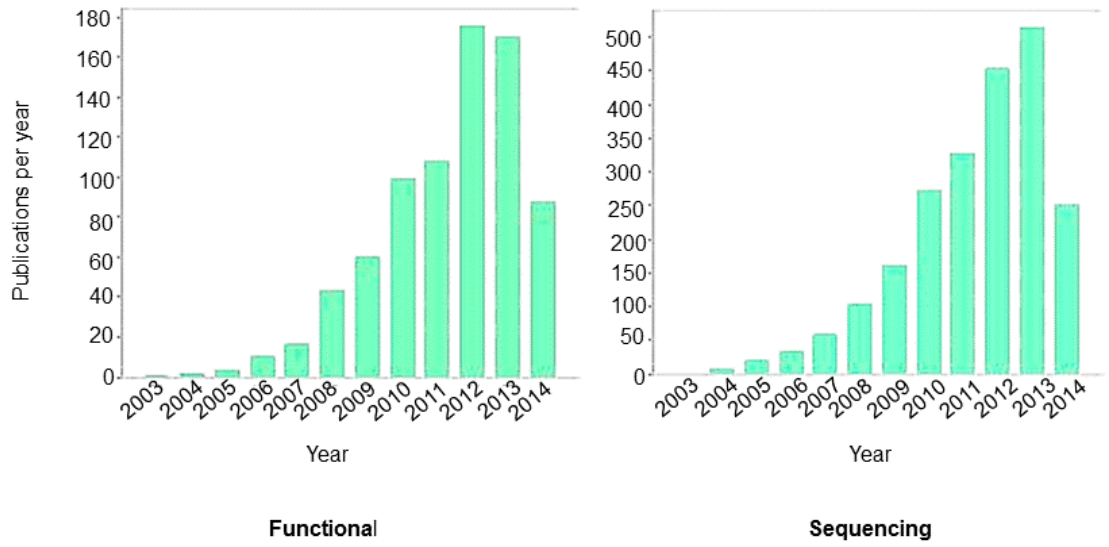


Figure 5.1. Number of publications over the last 10 years when using the ISI Web of Knowledge and carrying out the simply searches; "Functional Metagenomics" and "Sequencing Metagenomics". Note the different ‘y’ axis scale for each graph showing that there have been many at least twice as many publications on sequencing metagenomics compared to functional.

There are four broad areas to consider when striving for a successful metagenomic project and these are; DNA extraction method and insert size, choice of vector, the surrogate host used for heterologous expression of metagenomic DNA, and the

screening approach for the product of interest. DNA extraction methods to obtain high yield, highly pure and un-bias DNA from the environment in question is of tantamount importance for both FM and SM. Insert size is also dependent on the biological molecule in question and considerations for both have been discussed at depth in Chapter 1.0. While the choice of vector (plasmids, fosmids, cosmids and Bacterial Artificial chromosomes or BACs) is partly dictated by the insert size, it may also be dictated by the choice of surrogate host used in the study. While the majority of FM screens to date have implemented a single host system for expression of metagenomic DNA, work such as that by Ekkers and colleague suggest that the use of multi-host expression systems would be more efficient in the ability to express foreign DNA [13]. One of the most significant hindrance of FM is the necessity of the surrogate host to heterologously express metagenomic DNA originating from a plethora of microorgansims of extensive genetic diversity. The surrogate host, which is most commonly *E. coli,* may be unable to correctly express such DNA. The reason for which some host bacteria may not be able to express certain genes can be attributed to a number of factors, including different preferences for different codon usage during translation commonly referred to as the codon usage bias (CUB). The significance of this usage, in terms of FM, means that codons present in an inserted gene may not be recognised by the host organism and so will not become translated correctly and as a result, may not be expressed at all [14, 15]. Further to the problem of heterologous expression includes the necessity for the presence of cis regulatory elements that are non-coding regions of DNA that still have an essential role in DNA transcription such as promoter sequences which would result in no product being expressed and trans-acting elements such as signal peptides which would result in failure to secrete the product [13, 16]. Moreover, even if a surrogate host were able to successfully express the exogenous DNA, as the DNA is potentially originally from an organism of unknown taxonomy, there is likelihood that the gene product could infer a toxic effect on the host, directly killing the host or at least causing some form of growth inhibition hindering the screen.

Despite these difficulties, FM has been implemented to successfully isolate many hydrolytic enzymes (See Table 1.1 Chapter 1.0). There have been a number of solutions to the problems of heterologous expression of exogenous DNA in *E. coli*. For example, Leggweie et al., alleviated the problem of cis-acting factors by developing an inducible dual promoter transposon system termed MUExpress which would randomly insert into

metagenomic DNA and *E. coli* could express inserted genes regardless of their orientation within the vector [17].

However in comparison, FM projects isolating the group of hydrolytic enzymes that we are most interested in; proteases, for reasons described elsewhere (Chapter 1 section 1.3 and Chapter 7) remain relatively sparse (please see table 5.1 and compare to table 1.1 of Chapter 1 which shows enzymes isolated from most recent FM publications).

Table 5.2 Protease enzymes that have been isolated using functional metagenomics and if and how they were further validated so we can be certain that the protease phenotype observed was that of proteases and not other hydrolytic enzymes. 3 of the 9 studies failed to isolate any proteases. The majority of studies also implement skimmed milk agar as the screening means.

Protease Identified	Environment	Screening mechanism	Validated?	Reference
Alkaline serine protease	Goat skin surface metagenome	Skimmed milk agar	Zymography and functional characterisation by azocasein assay	[18]
Subtilisin-like serine protease	Red sea Atlantis-II brine pool	Skimmed milk agar	No further functional characterisation, pyrosequencing used to assess sequence	[19]
Oxidant-stable serine protease	Forest soil	AZCL-casein	Functional characterisation by azocasein assay	[20]
Serine protease x 2	Gobi and Death valley Deserts	Skimmed milk agar	Functional characterisation by azocasein assay	[21]
Metalloprotease x 2	Soil	Skimmed milk agar	Further screens on azoalbumin and azocasein assay	[22]
Subtilisin-like serine protease	Coastal sediment	Skimmed milk agar	Casein	[23]
Unknown type, 3 x protease clones	Antarctic soil	Skimmed milk agar	No	[24]
No protease phenotype observed	Soil	Skimmed milk agar	N/A	[25]
No protease phenotype observed	Human gut	Skimmed milk agar	N/A	[26]
No protease phenotype observed	Enriched Biomass	Skimmed milk agar	N/A	[27]

A number of research groups have aimed to isolate proteases using FM techniques and have failed at this endeavour; (Table 5.1[22, 25, 28]). Purohit and Singh [29] successfully isolated an alkaline protease from a metagenomic library created from a saline habitat. However they did not functionally screen their library. Instead, they designed degenerate primers from multiple alignments of the conserved region of other alkaline proteases from a species isolated from saline environments and conducted PCR to amplify target genes from the metagenome. Gene-targeting methods such as this are often implemented as an alternative to FM screening and are often successful [29-32]. However, this method requires a significant amount of *a priori* knowledge of sequence data in order to conduct an alignment of conserved regions as the basis for designing primers and therefore, rarely can non-homologous, novel genes be identified with these PCR-based methods.

Aside from the problems associated with heterologous gene expression that affect all FM screens, particularly the recognition of signal peptides which would be crucial for observation of a proteolytic phenotype, a further explanation for this insufficiency is likely due to the inherent role of proteases; degrading protein, in that although a host organism may very well be able to express an exogenous protease, this protease may exert detrimental effects on the host due to degradation of essential intracellular protein such as cell wall and cell membrane components as well as many other potential targets. It has been demonstrated before that certain bacterial proteases have a lytic effect on an alternative expression host [33]. Additionally, extracellular proteases often require forms of posttranslational modification in order to become active, increasing the genetic capacity required by the surrogate host for exogenous protease expression. Thus, this emphasises the necessity of developing functional metagenomic techniques that aim to ameliorate the results that have so far been achieved.

While *E. coli* has been shown capable of expressing bacterial genes of vast taxonomic diversity [9] most of these are Gram negatives. And these examples are from purified culture genome extractions not metagenomes. It has been suggested that exploration of broadening the surrogate host range beyond *E. coli,* and consequently expanding the genetic machinery for heterologous expression of foreign DNA should lead to a higher frequency of positive metagenomic screens [13, 34, 35]. For example Craig and colleagues constructed an FM library in *E. coli* and transferred to five different

proteobacteria as hosts for their FM screen. The library in each host was screened for altered pigmentation, altered colony morphologies and antibacterial activity. In total > 35 colonies displayed an altered pigmentation, 8 colonies showed antibiosis and 2 showed altered morphology, of these, *E. coli* accounted for 2 colonies conferring antibiosis and after retransformation, only one of these colonies recapitulated that phenotype [36]. Many of the compounds isolated were not active in the *E. coli* host demonstrating the usefulness of utilising alternative hosts. Martinez *et al.,* have also demonstrated that different hosts harbour different expression capabilities and also highlight that the environment under scrutiny is an important consideration when developing alternative host systems [35].

Studies such as those by Eckburg *et al.,* [37], Turnbaugh *et al.,* [38] and Andersson *et al.,* [39] have highlighted that Gram positive bacteria can make up a significant proportion in the human gut. Gram positive bacteria have different mechanisms for the secretion and activation of surface bound or extracellular protein [40] suggesting again that *E. coli* may not have the genetic capability to effectively express and secrete active proteases from the vastly Gram positive human gut.

Bacillus species' secretory pathways have been well characterised and have been used extensively in industry due to their high secretion capacity in the production of heterologous protein [41]. *Bacillus subtilis* has a GRAS (generally regarded as safe) status and is commonly used as a host for the large-scale production of industrial proteases and has a number of proteases itself, many of which are extracellular and therefore it has an effective and well-characterised means of protein secretion [42]. Also, it's ease of genetic manipulation means a number of tools have been developed for *B. subtilis*, particularly regarding the development of modified strains, transformation procedures, and the development of vectors for transformation. However, its extracellular proteases pose a problem for FM screens for two reasons. Firstly, they potentially will recognise and degrade heterologous proteins and secondly, secretion of endogenous proteases will inhibit the screen for exogenous metagenomics proteases. These obstacles have been alleviated by the construction of genetically modified strains that are protease deficient strains of *Bacillus subtilis* such as WB800, a strain engineered to be deficient in 6 extracellular protease genes [43] and WB800N, a strain with 8 null

extracellular protease genes [44]. These strains have also been specifically designed for use with secretion vectors.

A further problem includes that while many plasmids replicate stably in *B. subtilis* under normal conditions, upon insertion of recombinant DNA problems arise regarding structural and segregational stability. Instability is caused by the rolling circle mode of replication of plasmids with a single-stranded (ss) DNA intermediate [45]. During the ss DNA intermediate, short direct repeats and intervening DNA may often get deleted resulting in instability [46]. This problem can be ameliorated by the development of vectors with an expression cassette inserted into a non-essential gene such as the *lacA* encoding for β-galactosidase which allowed for controlled expression of inserted genes [47].

AIMS

We aim to further our understanding of proteases produced by the gut microbiota. One of the aims of this research will be to conduct a culture-dependent study to isolate protease producers from the cultivable gut microbiota to potentially compare to proteases isolated via culture-independent techniques. As determined in Chapter 4.0, it is clear that it is not suitable to solely rely on *E.coli* as a single surrogate host for functional metagenomics studies particularly when attempting to isolate proteases. Therefore, this research also mainly aims to explore new tools that facilitate the use of *E. coli* and *B. subtilis* as expression hosts for high-throughput FM library construction and screening specifically designed to isolate novel protease genes. These tools include the investigation of two strains of the alternative surrogate cloning host; *B. subtilis* for their use as hosts for FM screening for proteases. The most appropriate vector system was also to be explored, as well as determining optimum DNA insert sizes and fragmentation methods specifically for the expression of proteases.

Therefore the main aims of this research are:-

1. To conduct a culture-dependent analysis of proteolysis within the gut microbiota

2. Determine the usefulness of *B. subtilis* WB800N as an expression host for FM library screening for proteases.

3. Determine the most appropriate vectors for library construction.

4. Compare the robustness and efficiency of a Gram negative and Gram positive expression system using already characterised protease genes.

5. Determine the optimum DNA fragmentation methods and DNA fragment size for optimal protease capture from metagenomic libraries.

6. Implementation of the inferred novel FM tools to screen the human gut microbiota for novel proteases.

Different vectors will be tested for the suitability of this research. Also from the Bacillus Genetic Stock Centre (BGSC) we have four *E. coli-B. subtilis* shuttle vectors; pNZ8048, pNZ44, pHCMC04 and pHCMC05 [45] which can easily be transformed into *E. coli* for maintenance and can be transformed into WB800N to further examine the phenotype conferred by the recombinant vector.

5.2 RESULTS

5.2.1 *B. subtilis* WB800N adaptation for functional metagenomic screening for proteases

B. subtilis My2016 caused proteolytic cleavage of azo-dye from azocasein however, *B. subtilis* WB800N showed no activity (figure 5.2.1). Similarly, no proteolytic phenotype was observed when *B. subtilis* WB800N was cultured on Lactose-free skimmed milk agar, yet *B. subtilis* MY2016 showed a distinctive halo indicating proteolytic degradation of the media. These results show that it is the *nprE* and *nprB* neutral protease subtilisin, *epr* serine protease gene, bacillopeptidase F, *WprA* protease, and serine protease *Vpr* are responsible for extracellular proteolytic degradation in *Bacillus subtilis* although *nrpE* and *Apre* are responsible for approximately 70% of extracellular proteolytic activity and are the only two enzymes involved in protein degradation for nutritional purposes. The elimination of these protease gene undertaken by Wu et al., [43] whose research produced a strain of *B. subtilis*, with 8 proteases knocked out and with chromosomal resistance to chloramphenicol. This strain was further modified by Nguyen et al., [45] by insertion of a neomycin cassette into the chloramphenicol gene thus disrupting it rendering this new strain (B. subtilis WB800N) susceptible to chloramphenicol, but resistant to neomycin. This strain showed no observable proteolytic phenotype with the methods described. In addition, these protease genes were initially knocked out as they are known to recognise and degrade heterologous protein, the absence of such proteases and a proteolytic phenotype therefore makes *B. subtilis* WB800N suitable surrogate hosts for functional expression of proteases from the microbial metagenome. A number of different bacillus transformation procedures were implemented in order to determine the most efficient. The only transformation procedures that was successful was the MM competence medium method which is described in the general materials and methods section and the MD method (supplementary materials) and the transformation efficiencies are shown in table 5.1 alongside the transformation efficiencies of the *E. coli* hosts, when transformed with pNZ44 empty vector. The, MD medium method showed fewer colonies after transformation (table 5.1) and so was not used for future transformations. Most of the procedures resulting in no colonies after 12 and 24 hours of incubation although colonies did begin to develop after 48 hours, however, plasmid preparations of these colonies revealed no plasmid present and therefore it was presumed that these

transformation attempts had failed) these methods are detailed here in supplementary materials 5.5.1→5.5.4). Transformation efficiencies for *B. subtilis* WB800N still remained far lower than the transformation efficiencies in both strains of *E. coli.*

pNZ44, a vector originally designed by McGrath and colleagues [48] was the choice for use in this research due to the necessity for a low copy number plasmid to reduce gene dosage so as to be less toxic to the host, it has an origin of replication for both Gram negative and Gram positive organisms, it has a strong, constitutive promoter useful for optimal expression of recombinant protein [49]. The other plasmids were deemed too large for efficient tranformstion of large inserts and so were not used in this study.

Table 5.1 A comparison of transformation efficiencies between the protease deficient cloning host for this study; *B subtilis* WB800N with a commercially available chemically competent *E. coli* a commercially available electrocompetent *E. coli* that showed no positive phenotype on LF-SMA. [a]100 µl transformation solution was plated, [b] 200 µl was plated. Transformation efficiency was calculated using the following formula:

TE = (no. colonies/ pg DNA) x (1 x 10^6 pg/ µg) x (volume of transformants/volume plated) x dilution factor

The number highlighted in bold represent the transformation efficiency when using recombinant pNZ44.

Strain	Transformants/10^6 cells									
	MD				MM				Electroporation		Chemical	
	1[a]		2[b]		1		2					
***B. subtilis* WB800N**	0.32	0.29	0.76	0.26	0.38	0.56	2.28	1.5	n/a		n/a	
Chemically competent *E.coli*	n/a				n/a				n/a		6.75	3.2
Electro-competent *E.coli*	n/a				n/a				14.85	9.7	n/a	

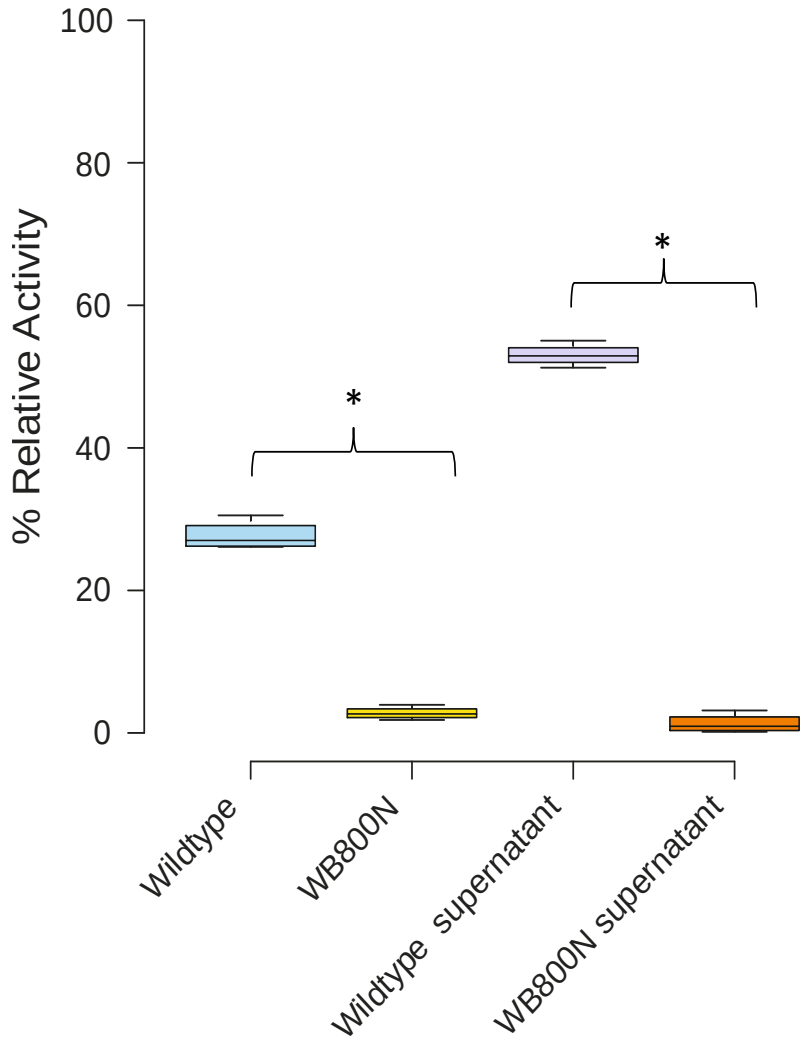


Figure 5.2.1 Protease activity assay using azo-casein as the substrate for wildtype B. subtilis MY16, and B. subtilis WB800N. The assay was also conducted using cell-free supernatant from both organisms. Significantly different results ($P = \leqq 0.05$) are represented with an '*'.

5.2.2 Culture-dependent Identification of Protease producing feacal bacteria

Serial dilutions of faecal samples from three individuals were plated on LFSMA and incubated under aerobic and anaerobic conditions as described in 5.4. All samples exhibited a large amount of proteolysis on the agar. Dilutions of 10^{-6} were only useful for identifying single colonies that were secreting proteases into the medium. Colonies exhibiting a clear zone of clearing after 24 hour incubation were streaked to purity on LFSMA to ensure the protease phenotype was recapitulated. Isolates that exhibited different colony morphology, colour and size were selected. 16S rRNA gene sequence analysis revealed all aerobic isolates to share greatest sequence identity to members of the *Enterococcus* genera, specifically; most of them shared ≥ 98 % sequence identity to *E. faecalis* except for one isolate which showed highest sequence identity to *Lactobacillus casei* (Table 5.2). The anaerobic isolates were shown to share ≥ 75% sequence identity to members of the genera; *Bacteroides* and the *Bifidobacteria* except for one isolate which shared 92% sequence identity with a previously uncultured microorganism; clone ELU 0138 (Table 5.2).

Table 5.2 The results of 16S rRNA gene analysis on protease producers isolated after aerobic and anaerobic culture on faecal sample dilutions from 3 individuals. Results of organisms identified with the greatest sequence similarity are shown.

Sample ID	16srRNA Gene ID	Identity (%)	Query Coverage (%)	E-value
Aerobic isolate 1	*Enterococcus faecalis*	99	100	0.0
Aerobic isolate 2	*Enterococcus faecalis*	99	100	0.0
Aerobic isolate 3	*Enterococcus faecalis*	99	100	0.0
Aerobic isolate 4	*Enterococcus* sp.	99	100	0.0
Aerobic isolate 5	*Enterococcus faecalis*	98	100	0.0
Aerobic isolate 6	*Enterococcus faecalis*	99	100	0.0
Aerobic isolate 7	*Enterococcus faecalis*	99	100	0.0
Aerobic isolate 8	*Lactobacillus casei*	97	100	0.0
Anaerobic Isolate 1	*Bacteroides fragilis*	85	100	0.0
Anaerobic Isolate 2	*Bifidobacterium bifidum*	93	97	0.0
Anaerobic Isolate 3	*Bifidobacteirum longum* subsp *longum*	97	92	4e-10
Anaerobic Isolate 4	*Bifidobacterium pseudcatenlatum*	75	92	2e-10
Anaerobic Isolate 5	Uncultured organism clone ELU0138	92	87	6e-14
Anaerobic Isolate 6	*Bacteroides thetaiotaomicron*	91	98	2e-13

5.2.3 Comparison of Expression of Characterised Proteases in a Gram negative and a Gram positive host

In order to test the hypothesis that the Gram positive host *B. subtilis* WB800N was better suited for expression of proteases from the gut metagenome, the ability of this strain to express characterised protease genes from both Gram negative and Gram positive organisms was tested against *E. coli.* The well characterised proteases; NprE (*Bacillus subtilis* MY2016), GelE (*Enterococcus faecalis* isolated from this aerobic study), LasB (P*seudomonas aeruginosa* PAO1) were used in this study. The organisms were first screening on LFSMA to ensure they exhibited a proteolytic phenotype. Cell-free supernatants were also plated to ensure the organisms were actively secreting proteases (figure 5.2.2). The protease genes were PCR amplified from DNA extracted from pure cultures of the relevant microorgansisms including flanking DNA regions and restriction enzyme sites (see materials and methods; primer information). After cloning into pNZ44 the plasmids were transformed into *B. subtilis* WB800N and *E. coli* XL1 blue supercompetent cells (Agilent, Berkshire) transformants were picked into 96 wells plates using a Flexys® colony picking robot (Genomic Solutions) and also screened on LFSMA for protease activity. *B. subtilis* WB800N with pNZ44-*gelE* recapitulated a proteolytic phenotype. The clones with a positive phenotype were streaked to purity and the plasmids were extracted upon recapitulation of the phenotype once again. PCR were performed using the primers deigned to amplify the multiple cloning site region of pNZ44 and amplified product was sequenced. Similarly, the purified plasmids were re-digested with an appropriate restriction enzyme, which does not cut in the inserted protease gene sequence and the putative protease product was sequenced using the appropriate primers for that protease gene. pNZ44-*B.subtilis-nprE* frame 1 shared 100% identity and 99% coverage with gelatinase E of numerous *E. faecalis* strains including *E. faecalis* 62, *E. faecalis* H81 and *E. faecalis* GM and shared 100% homology with the active site regions and metal ligands that are needed for activity (table 5.2 and figure 5.2.3). pNZ44-*E.coli-gelE* that did not exhibit a protease phenotype, but did share sequence similarity with sequenced *gelE* genes. However, for unknown reasons, the active site residues and/or the metal ligand binding sites both needed for protease activity were not present in the translated sequence that gave positive hits for *gelE* (figure 5.3).

B. subtilis WB800N also conferred a protease phenotype when transformed with pNZ44 containing the *nprE* neutral protease gene and once again *E. coli* did not. The sequences *E. coli nprE* sequences lacked homology in certain areas of the sequence (figure 5.2.3) and while the active site residue (HEXXH) was conserved, the metal ligand binding residue (E) was absent suggesting the protease could not bind metal ions for activity.

No positive phenotype was observed for any clones with pNZ44-LasB in either host. Surprisingly, both sequences shared homology regarding the active site residues and metal binding regions of the proteases potential mis-tranlsation may not be a cause for lack of protease expression.

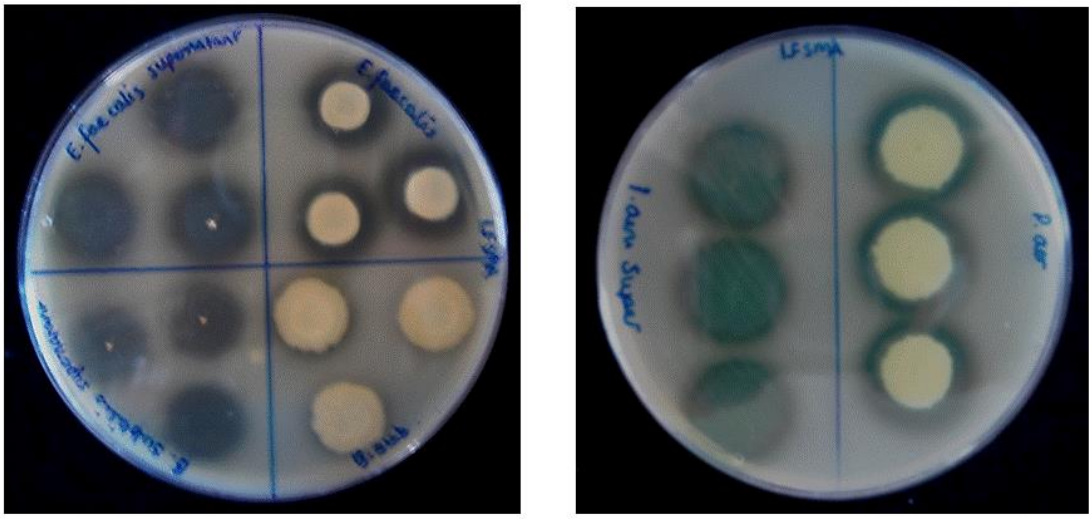


Figure 5.2.2 Protease activity of; *E. faecalis* (A; top right quarter) *E. faecalis* cell free supernatant (A top left quarter), *B. subtilis* MY2016 (A; bottom right quarter*) B. subtilis* MY2016 cell –free supernatant (A; bottom left quarter) and *P. aeruginosa* PAO1 (B; right) and *P. aeruginosa* PAO1 cell free supernatant (B; left).

Table 5.3 The results of homology searches of the deduced sequences of *gelE*, *lasB* and *nprE* after cloning in chemically competent E.coli and *B. subtilis* WB800N using NCBI BLASTn. The top 3 matches are shown along with the Accession number (in brackets), the percentage identity, E-value and the query coverage.

Sample	BLAST ID	Identity (%)	E-value	Query Coverage
GelE in *E. coli*	1.*E. faecalis* (*gelE*) complete cds (D85393.1)	99	0.0	97
	2. *E. faecalis* (*gelE*) complete cds (M37185.1)	98	0.0	97
	3.*E. faecalis* GM (*gelE*) complete cds EF105504.1)	98	0.0	97
GelE in B. subtilis WB800N	1.*E. facecalis* 62 (*gelE*) (CP002491.1)	99	0.0	99
	2. *E. faecalis* GM (*gelE*) complete cds (EF105504.1)	99	0.0	98
	3. *E. faecalis* H81 (*gelE*) partial cds (EU862241.3	99	0.0	98
LasB in *E. coli*	1.*P. aeruginosa* DNA complete genome, strain NCGM199 (Ap014622.1)	99	0.0	90

	2.*P. aeruginosa* NCGM2 S1 DNA complete genome AP012280.1	99	0.0	90
	3.*P. aeruginosa* A2 elastase precursor (*LasB*) gene complete cds (JQ235840.1)	98	0.0	89
LasB in *B. subtilis* WB800N	1.*P. aeruginosa* RP73 complete genome (CP006245.1)	99	0.0	90
	2.*P. aeruginosa* DH2 complete genome (CP006245.1)	99	0.0	90
	3.*P. aeruginosa* strain ATCC 25619 elastase gene, complete cds (JX040483.1)	99	0.0	90
NprE in E. coli	1.*B. subtilis* subsp. subtilis strain AG1839, complete genome (CP008698.1)	100	0.0	99
	2.*B. subtilis* neural protease gene complete cds (U30932.1)	100	0.0	99
	3.*B. subtilis* nprE gene encoding neutral protease (KP01985.1)	100	0.0	99
NprE in *B. subtilis* WB800N	1.*B. subtilis* subsp. *subtilis* strain AG1839, complete genome (CP008698.1)	100	0.0	99
	2.*B. subtilis* neural protease gene complete cds (U30932.1)	100	0.0	99
	3.*B. subtilis* nprE gene encoding neutral protease (KP01985.1)	100	0.0	99

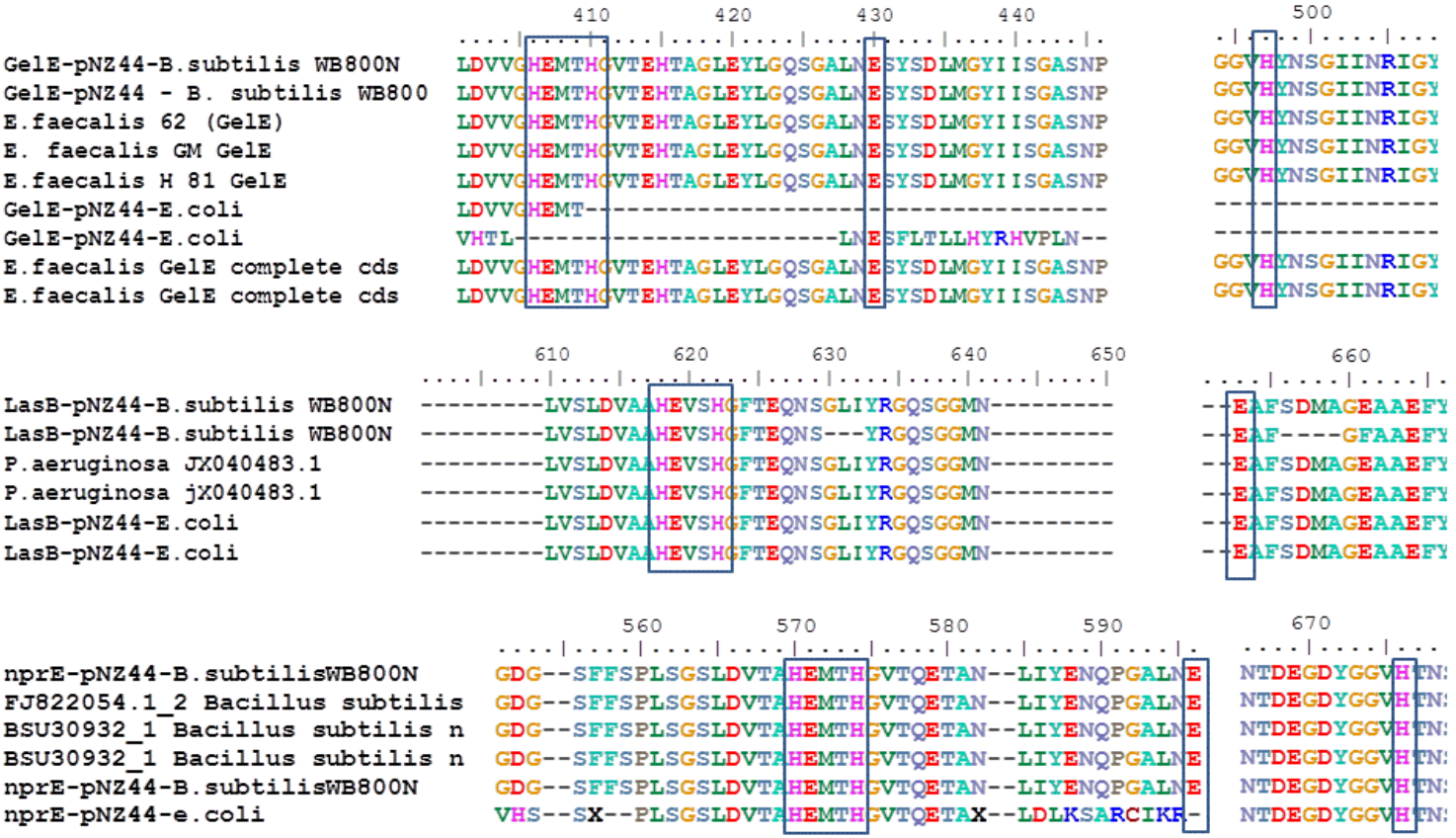


Figure 5.2.3 Alignments of the translated amino acid sequences of *gelE, lasB* and *nprE* along with their closets matches as determined by the NCBI Blastp program. Active site residues and metal ligand binding residues are outlined in order to determine whether missing regions could potentially be responsible for lack of expression of the protease gene.

5.2.4 Use of a Gram negative-Gram positive shuttle vector system to screen the human gut for proteases

The metagenomic library was constructed using genomic DNA extracted from the faeces of a healthy human volunteer. Faecal samples were collected and DNA extractions were extracted as described in the general materials and methods sections (Chapter 2.0). The genomic DNA was partially digested using *Fat*I as described in Chapter 2.0 and run on a gel (supplementary figure 5.5.5). The digested DNA of appropriate size was cloned into the digested pNZ44 plasmid which featured a chloramphenicol selection marker. The library was transformed into *E. coli and B. subtilis* WB800N. Transformants were selected by plating on TSA supplemented with chloramphenicol (5 μg/mL) and replica plated into 96-well plates using a colony picking robot (Genomic Solutions). A total of 19,200 clones were picked in *E. coli* and 9600 were picked in *B. subtilis* WB800N. Restriction digest analysis and sequencing of 10 clones from each library revealed the average insert size in *E. coli* to be ~ 994.9 bp giving an estimated library size of 19.1 Mb. In *B. subtilis* WB800N the average insert size was ~1103 bp giving an estimated library insert size of 10.5 Mb. *B. subtilis* WB800N remained particularly difficult to transform. Each library was screened initially on LFSMA for protease activity. However, no proteolytic clones were observed in any of the clones.

5.3 DISCUSSION

In the quest to unravel the vastly untapped resources of the microbial world, functional-based metagenomics remains currently the only potential means for unearthing completely novel compounds. In this research, our main interest lies with proteolytic enzymes isolated from bacteria for a range of reasons described elsewhere in this thesis (Chapter 1.0 and Chapter 7.0) particularly relating to the human gut. However, progress of FM is stinted by its dependence on the heterologous expression of foreign DNA and the numerous problems associated with this. While research into ameliorating the limitations of FM are well documented [17, 27, 50] even though the problem of an inefficient screening means for proteases has been improved [51] (Chapter 3.0) the problems associated with heterologous expression of proteases remain. In this research, we aimed to investigate an alternative host system for the expression of host proteases as a step towards accessing the degradome of the human gut microbiota. It was also deemed necessary to conduct our own culture-dependent analysis of the faecal degradome seeing as the overriding objective of this thesis was to further our knowledge of proteases in the human gut microbiota. Faecal samples from healthy individuals were also used to aerobically and anaerobically culture bacteria on lactose free skimmed milk agar as described in materials and methods and chapter 3.0. Colonies that exhibited proteolytic activity were streaked to purity, followed by DNA extraction, 16S rRNA gene amplification by PCR and sequencing this gene. All proteolytic isolates from aerobic cultures showed greatest sequence homology to *Enterococcus faecalis* (Table 5.2.2). *E. faecalis* is a facultative anaerobic bacterium that is commonly isolated from the human gastrointestinal tract [52, 53] and previous culture dependent studies on proteolytic faecal bacteria have isolated this microorganism too [54] so it is not surprising that here, *E. faecalis* was isolated aerobically. *E. faecalis* is usually considered a commensal or even a beneficial bacteria of the gut and certain strains have been used as probiotics for example one of the most well established is Symbioflor 1 (SymbioPharm, Herborn, Germany) which aids stimulation of the immune system to target recurring respiratory infections [55] and certain *E. faecalis* bacteriocins (e.g. the AS-48 bacteriocin of *E. faecalis* strain UGRA10) have been shown to exert antimicrobial activity against food-borne pathogens and spoilage agents such as *Listeria monocytogenes* [56].

However, on the contrary to these benefits, certain strains of *E. faecalis* are also recognised as significant pathogens implicated in disorders such as; bacteremia and endocarditis [57] and has been shown to induce inflammation in mouse models for IBD [58] and compromise epithelial barrier integrity [59] and strains often have multi-drug resistance [60]. Virulence of pathogen *E. faecalis* is commonly attributed to cytolysin, lipoteichoic acid aggregation substances [61], but also the activity of extracellular proteolytic enzymes, namely the zinc-dependent metalloprotease gelatinase (g*elE*) and a serine glutamyl protease (*sprE*) which have been implicated in IBD [62] and translocation across intestinal epithelia [63]. Research using *E. faecalis* mutants for the *gelE* and *sprE* genes have also shown to a significant reduction in protease activity [63] suggesting that it was one of these enzymes responsible for the protease phenotype used to isolate the organisms in this study.

The most numerous proteolytic bacteria identified following anaerobic culture were identified as *Bacteroides* spp. and *Bifidobacterium* spp. (Table 5.2.2b) following a BLAST homology search. This result is not surprising as *Bacteroides* spp. and *Bifiobacterium* spp. are well documented to be protease producers and are dominant members of the distal gut microbiota [64] and the *Bifidobacterium* species found in this studied have been isolated in other studies that aimed to isolate *bifidobacterial* species from human faecal samples before [65].

Similarly to the case of *E. faecalis*, *Bacteroides* spp. are often considered a mutual symbiont of the human gut, for example, the induced secretion of angiogenin by *Bacteroides thetaiotaomicron* [66], but they can also be the causative agents of significant infections. The phenomena of this transition from a commensal to a pathogenic phenotype may be caused by two factors; firstly, the vast genetic repertoire of the gut microbiota leads to extensive horizontal gene transfer of valuable genes that confer advantages to life in the gut. Secondly, organisms such as *E. faecalis* and *Bacteroides* spp. have vast genomic potential [67] [68] with different genes switched on to generate a pathogenic phenotype. Again, proteases of *Bacteroides* spp. are heavily implicated in pathogenesis, particularly in the species; *B. fragilis* of which the major enterotoxin is a zinc metalloprotease often referred as fragilysin which has been shown to be capable of breaking down intestinal brush border hydrolases [69] and degradation of E-cadherin, a transmembrane glycoprotein involved in epithelial cellular adhesion

[70] resulting in a loss in epithelial barrier integrity ('leaky' gut) and diarrhoea. Interestingly one isolate shared greatest sequence similarity to a previously uncultured microorganism [71] suggesting that perhaps this media, under anaerobic conditions was sufficient to support the growth of this potentially novel microorganism.

The results here further demonstrate that LFSMA is useful for culturing bacteria and screening for protease activity. Analysis of the MEROPS peptidase database, however, (http://merops.sanger.ac.uk/) [72] shows that some of these isolated organisms have hundreds of proteases so it is impossible to know which protease is responsible for the phenotype observed, and prior sequence knowledge would be needed to find out. Suggesting a metagenomic approach is important for identifying specific genes.

B. subtilis strains WB800 and WB800N were initially used to begin to develop metagenomic tools to implement in screening the gut microbiota for proteases. Neither strains inferred a proteolytic phenotype on LFSMA after 3 days of incubation and both showed significantly less activity that then wild type after azocasein assay analysis. Firstly, this suggests that the 6 protease genes knocked out were responsible for the protease phenotype observed by wild type *B. subtilis* and that both strains would be appropriate hosts for extracting proteases from the gut metagenome. While strain WB800 was determined to not be a consideration for interference with protease screens and would likely not affect heterologous protein, the strain contains chromosomal resistance to chloramphenicol, the selective resistance marker of all the shuttle vectors currently available for the organisms in this study. Whereas strain WB800N has a neomycin cassette inserted into the cat gene (conferring chloramphenicol resistance) thus rendering the strain resistant to neomycin but not chloramphenicol. Due to its compatibility with the vectors strain WB800N was consequently deemed more appropriate for this study and was used for all further experiments. However, no research has been previously carried out into the effect of the absence of such proteases on competence development and thus transformation efficiency. Initially, there was significant difficulty when trying to successfully transform WB800N, probably due to their large size and low copy number. The several transformations (see section5.4) that were attempted prior to estasblishemt of a successful protocol, helped to optimise the conditions for this new strain. Culture vessel size proved to be an important factor indicating that the strain required significant aeration when growing. The absence of

tryptophan had a great effect and needed to be included in the transformation medium. Several transformation methodologies were attempted in order to determine the greatest transformation efficiency. Transformations were tested using empty vector and mg(metagenomic) heterologous vector since the addition of small inserts (1-10kbp) into the bacteria were hypothesised to put a certain amount of stress and energy usage on the bacteria and potentially decrease transformation efficiencies. Transformation efficiencies of *B. subtilis* WB800N were compared with commercially available strains of *E.coli* (XL1-blue and Mega X DH10B™ $T1^R$ electrocompetent cells).

An external collaboration with the Laboratory of Professor Jeffrey Errington at Newcastle University, a laboratory with expertise in the genetics of *B. subtilis*, helped to determine the optimal transformation conditions for the genetically modified strain of *B. subtilis*. Growing *B. subtilis* WB800N in MM increased transformation efficiency 3 fold (Table 5.1) compared to using the MD media protocol when plating 200 µL of the transformation mix. Transformation efficiency decreased upon addition of heterologous plasmid approximately 1.5 fold when using the MM protocol and nearly 3- fold when using the MD protocol. It was found that 3 µL of plasmid or ligation reaction gave optimal efficiency when plating 200 µL (Table 5.1). It was also determined that mutant strains of *Bacillus subtilis* are particularly sensitive to temperature alterations when put under the 'stress' of competence induction via starvation media and transformation [73] . When all steps of the process were carried out at 37°C optimal transformation was observed. However, if the bacteria were subject to temperature deviations from 37 °C, very few to zero colonies were observed after incubation (table 5.1). Therefore all subsequent transformations were carried out in constant temperature rooms at 37°C using the MM protocol described previously (Chapter 2.0).

In order to test the hypothesis that the Gram positive host *B. subtilis* WB800N was better suited for expression of proteases from the gut metagenome, the ability of this strain to express characterised protease genes from both Gram negative and Gram positive organisms was tested against *E. coli.* The well characterised proteases; *nprE* (*Bacillus subtilis* strain 168), *gelE* (*Enterococcus faecalis* isolate from aerobic culture-dependent study previously conducted in this research), l*asB* (*Pseudomonas aeruginosa* PAO1) were PCR amplified from DNA extracted from pure cultures of the relevant

microorgansisms including flanking DNA regions and restriction enzyme sites (see materials and methods; primer information). After cloning into pNZ44 the plasmids were transformed into *B. subtilis* WB800N and *E. coli* XL1 blue supercompetent cells (Agilent, Berkshire) transformants were picked into 96 wells plates using a Flexys® colony picking robot (Genomic Solutions) and also screened on LFSMA for protease activity. *B. subtilis* WB800N with pNZ44-*nprE* recapitulated a proteolytic phenotype as it did for *pNZ44-gelE* but no positive phenotype was observed for *pNZ44-lasB. E. coli* Xl1-blue expressed no signs of protease activity on LFSMA for any of the protease genes. The heterologous expression of *gelE* has been attempted in *E. coli* before [74, 75] with similar unsuccessful results and it has been hypothesised that *E. coli* cannot recognise the Gram positive signal sequence for protein secretion. This theory is further supported by the evidence that *gelE* has been successfully heterologously expressed not only in *gelE* deficient strains of *E. faecalis,* but also in alternative Gram positive hosts such as *Lactococcus lactis, Streptococcus pyogenes* [75] and now from this study; *B. subtilis.* Following sequence analysis of the translated *gelE* sequence in both hosts It was revealed that in *B. subtilis* WB800N the active sites and metal binding regions needed for protease activity were conserved with the amino acid *gelE* sequence of numerous strains of *E. faecalis* however *E. coli* had failed to express this region of the gene which may be an explanation for the lack of expression observed.

Similarly, heterologous expression and secretion of the *nprE* gene has been successful in protease-deficient strains of *Bacillus stearothermophilus* and *B. subtilis* MT-1 [76] and now from this study, *B. subtilis* WB800N. Results from a previous study suggest that failure of *E. coli* to express *nprE* is due to the lytic effect of the gene product on the host cells [33] *NprE* has been successfully expressed in *E. coli,* but only upon removal of the ribosomal binding site [33] which would not be useful for a functional metagenomic screen. *NprE* has been successfully cloned into *Lacotococcus lactis* subsp. *lactis* [77] another Gram positive organism, eukaryotic microorganisms, namely *Saccharomyces cerevisiae,* have also been explored for the expression of proteases such as that encoded by *nprE*. However, it has been found that the eukaryotic host could not secrete *nprE* until manipulation of the sequence with a yeast invertase signal peptide [78] and although this resulted in successful secretion, the protein was still not active due to high glycosylation of the protein [78]. This suggests that perhaps eukaryotic hosts are not appropriate for FM screening of bacterial metagenomes specifically for protease activity

due to forms of post-translational modification. *NprE* in *E. coli* failed to express a protease phenotype and sequence analysis suggests that this may be due to the lacking translation of the metal binding regions of the amino acid sequence. This may also be a defence mechanism by the *E. coli* host to avoid self-lysis.

When transformed with pNZ44-*LasB* neither hosts demonstrated a protease phenotype despite apparent amino acid sequence homology (figure 5.3). *LasB* has been successfully expressed in *E. coli* in a previous study [79]. However, it was found that *LasB* expression is dependent on several genes including an autoinducer [80] and a *LasR* positive regulatory gene [81]. This is the most likely expression explains the problem with expression in both hosts in this study. Other studies have also observed that *E. coli* was unable to secrete the active elastase protease into the extracellular medium, but upon cell lysis, protease activity was observed [82]. High levels of post translational modifications were also witnessed when *lasB* was cloned into *Pichia pastoris,* however, in this instance, glycosylation did not inhibit protease activity [83] suggesting perhaps a eukaryotic host is appropriate for proteases that require numerous genes for expression and activity.

A metagenomic library was created in both surrogate hosts and screened for protease activity on LFSMA. It is estimated that proteases constitute 2-3% of the genomes of commensals [72] and ~ 5% of the genomes of pathogens [84]. Therefore large libraries are needed in order to 'capture' a protease gene. The isolation of proteases via FM techniques remains a challenge and to date, no proteases have been isolated from the human gut using these techniques. Unfortunately, this research was also unable to isolate proteases from the human gut microbiota. However, this research has highlighted some further issues and limitations with the techniques implemented which will significantly help to determine alternative strategies for improving screens in the future. Firstly, the fact that large libraries are needed for successful for FM screens is very apparent. This may go some way in explaining the lack of proteases isolated following screening on the FM library in this research. Especially considering in the few projects that have successfully isolated proteases that tens of thousands of clones were screened, for example Wichmann and colleagues [85] screened a library comprising over 6 million clones for antibiotic resistance genes and Jeon and [86] colleagues screened over 600,000 clones to isolate genes encoding lipase enzymes. Specifically for FM studies

that have successfully isolated proteases, over tens of thousands of clones had to be screened. For examples Pushpam and co-workers [18] screened 70,000 clones to isolate 1 protease clone and Waschowitz et al., [22] screened 80,000 clones to isolate two protease genes. Less than 30,000 clones, as is the case with this study, is likely not a large enough library to isolate proteases. The *B. subtilis* WB800N library in particular only harboured 9600 clones too and very few FM studies have ever isolated novel genes with this fewer clones. However, there was little that could be done with this factor as, although a successful transformation procedure was implemented the *B. subtilis* strain WB800N remained difficult to transform throughout the study, transformations often resulted in no colonies at all, and the transformation efficiency was very low. Other strains devoid of extracellular proteases may be more appropriate for FM screens for proteases circumventing the relatively inefficient transformation methods that had to be implemented here. For example, Nguygen and colleagues suggest *Brevibacillus brevis* as a suitable host for the expression of foreign proteins [45] and methodologies for expression of foreign protein have proven highly efficient [87]. However, the issue here cannot be definitively attributed to the hosts used for FM library construction. Instead, it is likely that the insert size of the library was just too small to capture protease genes which suggests that more efforts need to be implemented in order to ensure only large DNA fragments are ligated into the cloning vector. Introducing a size –dependent separation of DNA molecules procedure, such pulse field gel electrophoresis (PFGE) or size exclusion chromatography might ameliorate this issue.

As previously discussed in Chapter 4.0, enrichment techniques may also be implemented to enhance the number of bacteria with the target metabolic trait. Future directions for unearthing these elusive enzymes may include such a strategy, for example, using faecal samples from animals with a high protein diet or to use samples from a population that is known to be high in proteolysis. In a proceeding chapter (Chapter 7.0) it has been found that protease activity with the gut is significantly higher in those suffering from IBD as well as diarrhoeic IBS [88, 89]. An interesting study conducted by Rampelli et al., [90] also revealed that there is an age-related increase in proteolytic activity within the gut microbiota and that Centenarians had a gut microbiota significantly enriched for proteolytic function. It could be hypothesised therefore, that this population offers an attractive target for obtaining samples for future FM library construction to screen for proteases.

A final discussion point here would be that perhaps more success would be had with a more direct method for mining for proteases such as that conducted by Purohit and colleagues [29]. In this study, they were particularly interested in alkaline proteases of halophillic bacteria from saline habitats. They took a data mining strategy to PCR amplify proteases genes from DNA extracted from this habitat. The argument against such a strategy is the requirement of a significant amount of *a priori* knowledge in order to be able to develop degenerate primers and conduct PCR, and as a consequence, you can never get a gene that is truly novel. While this is true, FM techniques are also failing to isolate truly novel genes too, as described in Chapter 4.0. Purohit *et al.,* did manage to isolate a protease gene that exhibited different characteristics to the proteases used to generate the primers [29] suggesting the techniques may work for isolating proteases with novel functions.

5.4 MATERIALS AND METHODS

5.4.1 Culture-dependent analysis of protease-producing bacteria in the gut microbiota

In order to identify proteolytic bacteria within the gut microbiota that we are able to cultivate, aerobic and anaerobic conditions were utilised to facilitate the growth of as many proteolytic organisms as possible. Faecal samples (1 g wet weight) from three individuals were serially diluted in sterile Ringer's solution and dilutions were prepared up to 10^{-8}. Dilutions for the aerobic culture were plated in triplicate onto LFSMA and observed after 12 hours and 24 hours incubation at 37 °C. For anaerobic culture, all procedures were conducted in a Gallenkamp anaerobic cabinet containing 10 % H_2, 10 % CO_2 and 80% N_2 atmosphere. LFSMA plates were preincubated in the anaerobic cabinet for 48 hours prior to inoculation. Anaerobic plates were incubated in the anaerobic cabinet for 7 days. Production of secreted proteases was indicated by zones of clearing around colonies. These colonies were selected and streaked to purity under their appropriate conditions to ensure recapitulation of the protease phenotype. Upon validation, freezer stocks of each isolate were prepared as described in the general methods section (Chapter 2.0). A chelex PCR to amplify the 16S rRNA gene was also performed on each isolate as described in the general methods chapter (section 2.5.3). The presence of amplified 16S rRNA gene was verified by gel electrophoresis using a 1 % gel and visualisation with UV. Samples were purified using a QIAquick PCR clean-up kit (Qiagen) according to the manufacturer's instructions. The sequence of the16S rRNA from each isolate was carried out by Eurofins MWG Operon (Ebersberg, Germany) Sequences were compared against the BLAST database [91] and closest matches were recorded.

5.4.2 Amplification of *gelE*, *nprE* and *lasB* by PCR

In order to attempt to express characterised proteases from 3 microorganisms in *E. coli* and *B. subtilis* WB800N, the genes encoding the proteases were amplified by PCR. Firstly, each organism with the known protease (*P. aeruginosa* PAO1, *E. faecalis* isolate from this study and wild type *B. subtilis* MY16) were grown on appropriate agar plates (see table 2.1 Chapter 2.0). DNA extractions were then performed as described in chapter 2.0 using the appropriate liquid media for each organism. PCR was performed to amplify *gelE* from *E. faecalis*, *nprE* from *B. subtilis* and *lasB* from *P. aeruginosa*. The primers are shown in table 5.4 and include restriction enzyme sequences used for cloning into pNZ44.

Table 5.4. Primer sequences for amplification of *gelE*, *lasB* and *nprE*. The restriction enzyme sites are underlined and the particular enzyme is mentioned in brackets. The source of the primer is also mentioned although these sequences have different or no R.E sites.

Target	Primer name	Primer Sequence	Ref
gelE	gelE F	5′>GAACCATGGTAAGGAAGGAGTTAATTGTTTGATGAAG<3′ (NcoI)	[75]
gelE	gelE R	5′>CTTCTGCAGTTTCATTCATTGACCAGAACAGATTC<3′ (PstI)	[75]
lasB	lasB F	5′>CGCCATGGATGAAGAAGGTTTCTACGCTT<3 (NcoI)	[83]
lasB	lasB R	5′>GCGCATGCTTACAACGCGCTCGG<3 (SphI)	[83]
nprE	nprE F	5′>GCGGCTCCATGGCTCTTTATGCAA <3′ (NcoI)	[78]
nprE	nprE R	5′>CAAAATAAAGTAGCATGCGCCGCCGCCAC<3′(sphI)	[78]

The amplification of *gelE* was conducted as described by Waters et al., [75]. Following extraction of DNA the PCR was carried out in triplicate with the following conditions: initial denaturation at 94°C for 30 s, 55°C for 30 s, and 72°C for 2 min repeated 25 times; and then 72°C for 30 s. The products (~1.5 kb) were pooled and purified using the QIAquick purification kit (Qiagen) according to their instructions.

Following DNA extraction from *P. aeruginosa* PAO1, the *lasB* was amplified based on the reaction outlined by Han and colleagues [83] the PCR was carried out in triplicate under the following conditions; initial denaturing at 96°C for 5 min; subsequent 25 cycles of denaturing at 96°C for 40 s, annealing at 58°C for 30 s, and extension at 74°C for 2 min; and a final extension at 72°C for 5 min. The PCR products were analysed on a 1% (w/v) agarose gel by electrophoresis for the correct target band of ~1.5 kb in length [83]. The triplicate samples were again pooled and purified in preparation for the succeeding cloning steps.

To amplify the *nprE* gene from the DNA extracted from *B. subtilis* MY2016 the PCR was also carried out in triplicate using the following conditions; 95°C for 2 min followed by 25 cycles of 94°C for 30s, 52°C for 30s, 72°C for 2 min and a final extension step of 72°C for 5 min. PCR products (~1.5 kb) were pooled and purified.

Reactions for the amplification of all PCR products consisted of the same volumes and concentrations of PCR reagents as described in Chapter 2.0 (Section 2.5.3).

5.4.3 Cloning and screening of *gelE*, *nprE* and *lasB* into pNZ44

Following purification of PCR product, restriction digests with the appropriate enzyme (table 5.4) were performed on both the PCR product and purified pNZ44 to generate complementary sticky ends for efficient cloning. Plasmid and PCR product were purified once again and ligated at a 3:1 insert to vector ratio using T4 DNA ligase (Promega, Southampton, UK) according to the manufacturer's instructions except with an overnight incubation at 14 °C for ligation to occur. The ligation mixture (5 µl) was used to transform *E. coli* XL1 blue and *B. subtilis* WB800N according to the transformation procedures described in Chapter 2.0 (section 2.7). The transformation mix (200 µl per plate) was plated onto TSA supplemented with the appropriate antibiotic and incubated overnight at 37 °C. Transformants were picked using a Flexys® colony picking robot (Genomic Solutions, Cambridgeshire, UK) or by hand if there were few transformants using sterile toothpicks into freshly prepared freezer broth (per Litre: 10 g tryptone, 5g yeast extract, 10g NaCl, 44 mL ultrapure glycerol, 1.8g KH_2PO_4, 6.7g K_2HPO_4, 0.5g Sodium Citrate, 0.9g $(NH_4)_2SO_4$ adjust to pH 7.0 using 5N

NaOH and adjust the total volume to 1 L using dH_2O. Sterilise by autoclaving and after cooling, add 400 µl 1M $MgSO_4$) in 96-well plates. Duplicates of each plate were also prepared. The 96-well plates were used to inoculate LFSMA and following incubation, successful heterologous expression of the protease genes was indicated by the formation of clear halos around the colony. Triplicate positives and triplicate negative colonies were isolated and the plasmids were purified and restriction digests were performed to confirm an insert. Inserts were sub-cloned into the pGEM ®T-Easy vector and sequenced. Sequence analysis was conducted using NCBI BLASTn and BLASTp suites [91].

5.4.4 Metagenomic library construction and screening

Faecal samples were collected (Chapter 2.0 section 2.3.1) and genomic DNA was extracted from the samples as described in Chapter 2,0 (section 2.4.2). Extracted DNA was checked by running on a 1 % (w/v) agarose gel supplemented with 1 µg/mL safeview. DNA was subject to partial digest as described in chapter 2.0 (Section 2.5.1) (Supplementary figure 5.1) and DNA of the appropriate size (3-10 kb) (supplementary materials figure 5.2) was extracted using a QIAquick gel extraction kit (Qiagen) according to the manufacturers' instructions. pNZ44 was extracted as previously described (Chapter 2.0, 2.2.1) and digested (2.5.2). Digested plasmid (supplementary figure 5.2) was purified once again and ligated with digested gDNA (1-10 kb) at a 3:1 insert to vector ratio using T4 DNA ligase (Promega) according to the manufacturer's instructions except with an overnight incubation at 14 °C for ligation to occur. The ligation mixture (5 µl) was used to transform *E. coli* XL1 blue and *B. subtilis* WB800N according to the transformation procedures described in Chapter 2.0 (section 2.7). The transformation mix (200 µl per plate) was plated onto TSA supplemented with the appropriate antibiotic and incubated overnight at 37 °C. Transformants were picked using a Flexys® robot (Genomic solutions) or by hand into freshly prepared freezer in 96-well plates. Duplicates of each plate were also prepared. The 96-well plates were used to inoculate LFSMA to screen for protease activity.

5.5 REFERENCES

1. Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. Genome Biol *3*.
2. Staley, J.T., and Konopka, A. (1985). Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. Annual Review of Microbiology *39*, 321-346.
3. Ratjen, F. (2001). Changes in strategies for optimal antibacterial therapy in cystic fibrosis. International Journal of Antimicrobial Agents *17*, 93-96.
4. Rogers, G.B., Carroll, M.P., Serisier, D.J., Hockey, P.M., Jones, G., and Bruce, K.D. (2004). Characterization of Bacterial Community Diversity in Cystic Fibrosis Lung Infections by Use of 16S Ribosomal DNA Terminal Restriction Fragment Length Polymorphism Profiling. Journal of Clinical Microbiology *42*, 5176-5183.
5. Tunney, M.M., Field, T.R., Moriarty, T.F., Patrick, S., Doering, G., Muhlebach, M.S., Wolfgang, M.C., Boucher, R., Gilpin, D.F., McDowell, A., et al. (2008). Detection of Anaerobic Bacteria in High Numbers in Sputum from Patients with Cystic Fibrosis. American Journal of Respiratory and Critical Care Medicine *177*, 995-1001.
6. Finegold, S.M., Attebery, H.R., and Sutter, V.L. (1974). Effect of diet on human fecal flora: comparison of Japanese and American diets. The American Journal of Clinical Nutrition *27*, 1456-1469.
7. Ley, R.E., Peterson, D.A., and Gordon, J.I. (2006). Ecological and Evolutionary Forces Shaping Microbial Diversity in the Human Intestine. Cell *124*, 837-848.
8. Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., and Goodman, R.M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & Biology *5*, R245-R249.
9. Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiol. Mol. Biol. Rev. *68*, 669-685.
10. Courtois, S., Cappellano, C.M., Ball, M., Francou, F.-X., Normand, P., Helynck, G., Martinez, A., Kolvek, S.J., Hopke, J., Osburne, M.S., et al. (2003). Recombinant Environmental Libraries Provide Access to Microbial Diversity for Drug Discovery from Natural Products. Applied Environmental Microbiology. *69*, 49-55.
11. Thompson, S.A., Maani, E.V., Lindell, A.H., King, C.J., and McArthur, J.V. (2007). Novel Tetracycline Resistance Determinant Isolated from an Environmental Strain of Serratia marcescens. Applied Environmental Microbiology. *73*, 2199-2206.
12. Allen, H.K., Moe, L.A., Rodbumrer, J., Gaarder, A., and Handelsman, J. (2008). Functional metagenomics reveals diverse [beta]-lactamases in a remote Alaskan soil. ISME J *3*, 243-251.
13. Ekkers, D.M., Cretoiu, M.S., Kielak, A.M., and Elsas, J.D. (2012). "The great screen anomaly"-a new frontier in product discovery through functional metagenomics. Applied Microbiology Biotechnology *93*, 1005-1020.
14. Behura, S.K., and Severson, D.W.(2013). Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. Biological Reviews *88*, 49-61.

15. Kudla, G., Murray, A.W., Tollervey, D., and Plotkin, J.B. (2009). Coding-Sequence Determinants of Gene Expression in Escherichia coli. Science *324*, 255-258.
16. Angelov, A., Mientus, M., Liebl, S., and Liebl, W. (2009). A two-host fosmid system for functional screening of (meta)genomic libraries from extreme thermophiles. Systematic and Applied Microbiology *32*, 177-185.
17. Leggewie, C., Henning, H., Schmeisser, C., Streit, W.R., and Jaeger, K.-E. (2006). A novel transposon for functional expression of DNA libraries. Journal of Biotechnology *123*, 281-287.
18. Pushpam, L., Rajesh, T., and Gunasekaran, P.(2011). Identification and characterization of alkaline serine protease from goat skin surface metagenome. AMB Express *1*, 1-10.
19. El Said, N.H. (2012). Identification of potential subtilisin-like serine proteases from the microbial community of the Red Sea Atlantis-II brine pool using a metagenomic approach. In Department of Biology, Volume Biotechnology. (American University of Cairo).
20. Sophie Biver, D.P.a.M.V. (2013). Characterization of a new oxidant-stable serine protease isolated by functional metagenomics. SpringerPlus *2*.
21. Neveu, J., Regeard, C., and DuBow, M. (2011). Isolation and characterization of two serine proteases from metagenomic libraries of the Gobi and Death Valley deserts. Applied Microbiology and Biotechnology *91*, 635-644.
22. Waschkowitz, T., Rockstroh, S., and Daniel, R. (2009). Isolation and Characterization of Metalloproteases with a Novel Domain Structure by Construction and Screening of Metagenomic Libraries. Applied Environmental. Microbiology *75*, 2506-2516.
23. Zhang, Y., Zhao, J., and Zeng, R. (2011). Expression and characterization of a novel mesophilic protease from metagenomic library derived from Antarctic coastal sediment. Extremophiles *15*, 23-29.
24. Berlemont, R., Pipers, D., Delsaute, M., Angiono, F., Feller, G., Galleni, M., and Power, P.(2011). Exploring the Antarctic soil metagenome as a source of novel cold-adapted enzymes and genetic mobile elements. Reviews Argent Microbiology *43*, 94-103.
25. Rondon, M.R., August, P.R., Bettermann, A.D., Brady, S.F., Grossman, T.H., Liles, M.R., Loiacono, K.A., Lynch, B.A., MacNeil, I.A., Minor, C., et al. (2000). Cloning the Soil Metagenome: a Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms. Applied Environmental Microbiology *66*, 2541-2547.
26. Cecchini, D.A., Laville, E., Laguerre, S., Robe, P., Leclerc, M., Doré, J.l., Henrissat, B., Remaud-Siméon, M., Monsan, P., and Potocki-Véronèse, G. (2013). Functional Metagenomics Reveals Novel Pathways of Prebiotic Breakdown by Human Gut Bacteria. PLoS ONE *8*, e72766.
27. Lussier, F.-X., Chambenoit, O., Côté, A., Hupé, J.-F., Denis, F., Juteau, P., Beaudet, R., and Shareck, F. (2010). Construction and functional screening of a metagenomic library using a T7 RNA polymerase-based expression cosmid vector. Journal of Industrial Microbiology and Biotechnology, 1-8.
28. Jones, B.V., Sun, F., and Marchesi, J.R. (2007). Using skimmed milk agar to functionally screen a gut metagenomic library for proteases may lead to false positives. Letters in Applied Microbiology *45*, 418-420.

29. Purohit, M.K., and Singh, S.P.(2013). A metagenomic alkaline protease from saline habitat: Cloning, over-expression and functional attributes. International Journal of Biological Macromolecules *53*, 138-143.
30. Iwai, S., Chai, B., Sul, W.J., Cole, J.R., Hashsham, S.A., and Tiedje, J.M. (2009). Gene-targeted-metagenomics reveals extensive diversity of aromatic dioxygenase genes in the environment. ISME J *4*, 279-285.
31. Liebeton, K., and Eck, J. (2004). Identification and Expression in *E. coli* of Novel Nitrile Hydratases from the Metagenome. Engineering in Life Sciences *4*, 557-562.
32. Kotik, M. (2009). Novel genes retrieved from environmental DNA by polymerase chain reaction: Current genome-walking techniques for future metagenome applications. Journal of Biotechnology *144*, 75-82.
33. Wang, L.F., Ekkel, S.M., and Devenish, R.J. (1990). Expression in Escherichia coli of the Bacillus subtilis neutral protease gene (NPRE) lacking its ribosome binding site. Biochemistry international *22*, 1085-1093.
34. Wexler, M., and Johnston, A.W.B. (2010). Wide host-range cloning for functional metagenomics. Methods in Molecular Biology *668*, 77-96.
35. Martinez, A., Kolvek, S.J., Yip, C.L.T., Hopke, J., Brown, K.A., MacNeil, I.A., and Osburne, M.S. (2004). Genetically Modified Bacterial Strains and Novel Bacterial Artificial Chromosome Shuttle Vectors for Constructing Environmental Libraries and Detecting Heterologous Natural Products in Multiple Expression Hosts. Applied and Environmental Microbiology *70*, 2452-2463.
36. Craig, J.W., Chang, F.-Y., Kim, J.H., Obiajulu, S.C., and Brady, S.F.(2010). Expanding Small-Molecule Functional Metagenomics through Parallel Screening of Broad-Host-Range Cosmid Environmental DNA Libraries in Diverse Proteobacteria. Applied and Environmental Microbiology *76*, 1633-1641.
37. Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S.R., Nelson, K.E., and Relman, D.A. (2005). Diversity of the human intestinal microbial flora. Science. 2005 Jun 10;308(5728):1635-8. Epub 2005 Apr 14.
38. Turnbaugh, P.J., Quince, C., Faith, J.J., McHardy, A.C., Yatsunenko, T., Niazi, F., Affourtit, J., Egholm, M., Henrissat, B., Knight, R., et al. (2010). Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. Proceedinngs of the National Academy of Science U S A *107*, 7503-7508.
39. Andersson, A.F., Lindberg, M., Jakobsson, H., Bäckhed, F., Nyrén, P., and Engstrand, L. (2008). Comparative Analysis of Human Gut Microbiota by Barcoded Pyrosequencing. PLoS ONE *3*, e2836.
40. Schneewind, O., and Missiakas, D.M. (2012). Protein secretion and surface display in Gram-positive bacteria. Philosophical transactions of the Royal Society of London. Series B, Biological sciences *367*, 1123-1139.
41. Simonen, M., and Palva, I. (1993). Protein secretion in Bacillus species. Microbiological Reviews *57*, 109-137.
42. Lee, S.J., Pan, J.G., Park, S.H., and Choi, S.K. (2010). Development of a stationary phase-specific autoinducible expression system in *Bacillus subtilis*. Journal of Biotechnology *149*, 16-20.

43. Wu, X.C., Lee, W., Tran, L., and Wong, S.L. (1991). Engineering a *Bacillus subtilis* expression-secretion system with a strain deficient in six extracellular proteases. Journal of Bacteriology *173*, 4952-4958.
44. Wu, S.C., and Wong, S.L. (2002). Engineering of a *Bacillus subtilis* strain with adjustable levels of intracellular biotin for secretory production of functional streptavidin. Applied and Environmental Microbiology *68*, 1102-1108.
45. Nguyen, H.D., Nguyen, Q.A., Ferreira, R.C., Ferreira, L.C.S., Tran, L.T., and Schumann, W. (2005). Construction of plasmid-based expression vectors for *Bacillus subtilis* exhibiting full structural stability. Plasmid *54*, 241-248.
46. Bron, S., Holsappel, S., Venema, G., and Peeters, B.P. (1991). Plasmid deletion formation between short direct repeats in Bacillus subtilis is stimulated by single-stranded rolling-circle replication intermediates. Molecular & general genetics : MGG *226*, 88-96.
47. Hartl, B., Wehrl, W., Wiegert, T., Homuth, G., and Schumann, W. (2001). Development of a new integration site within the *Bacillus subtilis* chromosome and construction of compatible expression cassettes. Journal of Bacteriology *183*, 2696-2699.
48. McGrath, S., Fitzgerald, G.F., and van Sinderen, D. (2001). Improvement and optimization of two engineered phage resistance mechanisms in *Lactococcus lactis.* Appl Environ Microbiol. 2001 Feb;67(2):608-16.
49. van der Vossen, J.M., van der Lelie, D., and Venema, G. (1987). Isolation and characterization of *Streptococcus cremoris* Wg2-specific promoters. Applied and Environmental Microbiology *53*, 2452-2457.
50. Bernstein, J.R., Bulter, T., Shen, C.R., and Liao, J.C. (2007). Directed evolution of ribosomal protein S1 for enhanced translational efficiency of high GC *Rhodopseudomonas palustris* DNA in *Escherichia coli*. In Journal of Biological Chemistry, Volume 282. (United States), pp. 18929-18936.
51. Morris, L.S., Evans, J., and Marchesi, J.R. (2012). A robust plate assay for detection of extracellular microbial protease activity in metagenomic screens and pure cultures. Journal of Microbiological Methods *91*, 144-146.
52. Noble, C.J. (1978). Carriage of group D streptococci in the human bowel. Journal of Clinical Pathology *31*, 1182-1186.
53. Kayser, F.H. (2003). Safety aspects of enterococci from the medical point of view. International Jounral of Food Microbiology *88*, 255-262.
54. Macfarlane, G.T., Allison, C., Gibson, S.A., and Cummings, J.H. (1988). Contribution of the microflora to proteolysis in the human large intestine. The Journal of applied bacteriology *64*, 37-46.
55. Franz, C.M.A.P., Huch, M., Abriouel, H., Holzapfel, W., and Gálvez, A. (2011). Enterococci as probiotics and their implications in food safety. International Journal of Food Microbiology *151*, 125-140.
56. Cebrián, R., Baños, A., Valdivia, E., Pérez-Pulido, R., Martínez-Bueno, M., and Maqueda, M. (2012). Characterization of functional, safety, and probiotic properties of *Enterococcus faecalis* UGRA10, a new AS-48-producer strain. Food Microbiology *30*, 59-67.
57. Kayaoglu, G., and Ørstavik, D. (2004). Virulence Factors of *Enterococcus faecalis*: Relationship to Endodontic Disease. Critical Reviews in Oral Biology & Medicine *15*, 308-320.

58. Balish, E., and Warner, T. (2002). *Enterococcus faecalis* induces inflammatory bowel disease in interleukin-10 knockout mice. The American journal of pathology *160*, 2253-2257.
59. Steck, N., Hoffmann, M., Sava, I.G., Kim, S.C., Hahne, H., Tonkonogy, S.L., Mair, K., Krueger, D., Pruteanu, M., Shanahan, F., et al. (2011). *Enterococcus faecalis* Metalloprotease Compromises Epithelial Barrier and Contributes to Intestinal Inflammation. Gastroenterology *141*, 959-971.
60. Uttley, A.C., Collins, C.H., Naidoo, J., and George, R.C. (1988). Vancomycin-resistant enterococci. The Lancet *331*, 57-58.
61. Jett, B.D., Huycke, M.M., and Gilmore, M.S. (1994). Virulence of enterococci. Clinical Microbiology Reviews *7*, 462-478.
62. Steck, N., Mueller, K., Schemann, M., and Haller, D. (2013). Bacterial proteases in IBD and IBS. Gut.
63. Zeng, J., Teng, F., and Murray, B.E. (2005). Gelatinase is important for translocation of *Enterococcus faecalis* across polarized human enterocyte-like T84 cells. Infection and Immunity *73*, 1606-1612.
64. Bäckhed, F., Ley, R.E., Sonnenburg, J.L., Peterson, D.A., and Gordon, J.I. (2005). Host-Bacterial Mutualism in the Human Intestine. Science *307*, 1915-1920.
65. Turroni, F., Foroni, E., Pizzetti, P., Giubellini, V., Ribbera, A., Merusi, P., Cagnasso, P., Bizzarri, B., de'Angelis, G.L., Shanahan, F., et al. (2009). Exploring the Diversity of the Bifidobacterial Population in the Human Intestinal Tract. Applied and Environmental Microbiology *75*, 1534-1545.
66. Hooper, L.V., Stappenbeck, T.S., Hong, C.V., and Gordon, J.I. (2003). Angiogenins: a new class of microbicidal proteins involved in innate immunity. Nature Immunology *4*, 269-273.
67. Wexler, H.M. (2007). Bacteroides: the Good, the Bad, and the Nitty-Gritty. Clinical Microbiology Reviews *20*, 593-621.
68. Ó Cuív, P., Klaassens, E.S., Smith, W.J., Mondot, S., Durkin, A.S., Harkins, D.M., Foster, L., McCorrison, J., Torralba, M., Nelson, K.E., et al. (2013). Draft Genome Sequence of *Enterococcus faecalis* PC1.1, a Candidate Probiotic Strain Isolated from Human Feces. Genome Announcements *1*.
69. Riepe, S.P., Goldstein, J., and Alpers, D.H. (1980). Effect of secreted Bacteroides proteases on human intestinal brush border hydrolases. Journal of Clinical Investigation *66*, 314-322.
70. Wu, S., Lim, K.-C., Huang, J., Saidi, R.F., and Sears, C.L. (1998). *Bacteroides fragilis* enterotoxin cleaves the zonula adherens protein, E-cadherin. Proceedings of the National Academy of Sciences *95*, 14979-14984.
71. Li, E., Hamm, C.M., Gulati, A.S., Sartor, R.B., Chen, H., Wu, X., Zhang, T., Rohlf, F.J., Zhu, W., Gu, C., et al. (2012). Inflammatory bowel diseases phenotype, *C. difficile* and NOD2 genotype are associated with shifts in human ileum associated microbial composition. PLoS ONE *7*, e26284.
72. Rawlings, N.D., Barrett, A.J., and Bateman, A. (2014). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Research *40*, D343-D350.
73. Nijland, R., Burgess, J.G., Errington, J., and Veening, J.-W. (2010). Transformation of Environmental *Bacillus subtilis* Isolates by Transiently Inducing Genetic Competence. PLoS ONE *5*, e9724.

74. Su, Y.A., Sulavik, M.C., He, P., Makinen, K.K., Makinen, P.L., Fiedler, S., Wirth, R., and Clewell, D.B. (1991). Nucleotide sequence of the gelatinase gene (*gelE*) from *Enterococcus faecalis* subsp. *liquefaciens*. Infection and Immunity *59*, 415-420.
75. Waters, C.M., Antiporta, M.H., Murray, B.E., and Dunny, G.M. (2003). Role of the *Enterococcus faecalis GelE* Protease in Determination of Cellular Chain Length, Supernatant Pheromone Levels, and Degradation of Fibrin and Misfolded Surface Proteins. Journal of Bacteriology *185*, 3613-3623.
76. Fujii, M., Takagi, M., Imanaka, T., and Aiba, S. (1983). Molecular cloning of a thermostable neutral protease gene from *Bacillus stearothermophilus* in a vector plasmid and its expression in *Bacillus stearothermophilus* and *Bacillus subtilis*. Journal of Bacteriology *154*, 831-837.
77. van de Guchte, M., Kodde, J., van der Vossen, J.M., Kok, J., and Venema, G. (1990). Heterologous gene expression in *Lactococcus lactis* subsp. *lactis*: synthesis, secretion, and processing of the *Bacillus subtilis* neutral protease. Applied and environmental microbiology *56*, 2606-2611.
78. Wang, L.F., and Devenish, R.J. (1993). Expression of *Bacillus subtilis* neutral protease gene (*nprE*) in *Saccharomyces cerevisiae*. Journal of General Microbiology *139*, 343-347.
79. Bever, R.A., and Iglewski, B.H. (1988). Molecular characterization and nucleotide sequence of the *Pseudomonas aeruginosa* elastase structural gene. Journal of Bacteriology *170*, 4309-4314.
80. Passador, L., Cook, J., Gambello, M., Rust, L., and Iglewski, B. (1993). Expression of *Pseudomonas aeruginosa* virulence genes requires cell-to-cell communication. Science *260*, 1127-1130.
81. Gambello, M.J., and Iglewski, B.H. (1991). Cloning and characterization of the *Pseudomonas aeruginosa lasR* gene, a transcriptional activator of elastase expression. Journal of Bacteriology *173*, 3000-3009.
82. Ogino, H., Yokoo, J., Watanabe, F., and Ishikawa, H. (2000). Cloning and sequencing of a gene of organic solvent-stable protease secreted from *Pseudomonas aeruginosa* PST-01 and its expression in *Escherichia coli.* Biochemical engineering journal *5*, 191-200.
83. Han, M., Ding, H., Wang, J., Jin, M., and Yu, X. (2013). Expression of the *lasB* gene encoding an organic solvent-stable elastase in *Pichia pastoris* and potential applications of the recombinant enzymes in peptide synthesis. Biochemical engineering journal *77*, 154-160.
84. Maeda, H., and Molla, A. (1989). Pathogenic potentials of bacterial proteases. Clinica Chimica Acta *185*, 357-367.
85. Wichmann, F., Udikovic-Kolic, N., Andrew, S., and Handelsman, J. (2014). Diverse Antibiotic Resistance Genes in Dairy Cow Manure. mBio *5*.
86. Jeon, J.H., Kim, J.T., Lee, H.S., Kim, S.-J., Kang, S.G., Choi, S.H., and Lee, J.-H. (2011). Novel Lipolytic Enzymes Identified from Metagenomic Library of Deep-Sea Sediment. Evidence-Based Complementary and Alternative Medicine *2011*, 9.
87. Mizukami, M., Hanagata, H., and Miyauchi, A. (2010). Brevibacillus Expression System: Host-Vector System for Efficient Production of Secretory Proteins. Current Pharmaceutical Biotechnology *11*, 251-258.
88. Tooth, D., Garsed, K., Singh, G., Marciani, L., Lam, C., Fordham, I., Fields, A., Banwait, R., Lingaya, M., Layfield, R., et al. (2014). Characterisation of faecal

protease activity in irritable bowel syndrome with diarrhoea: origin and effect of gut transit. Gut *63*, 753-760.
89. Gecse, K., Róka, R., Ferrier, L., Leveque, M., Eutamene, H., Cartier, C., Ait-Belgnaoui, A., Rosztóczy, A., Izbéki, F., Fioramonti, J., et al. (2008). Increased faecal serine protease activity in diarrhoeic IBS patients: a colonic lumenal factor impairing colonic permeability and sensitivity. Gut *57*, 591-599.
90. Rampelli, S., Candela, M., Turroni, S., Biagi, E., Collino, S., Franceschi, C., O'Toole, P.W., and Brigidi, P. (2013). Functional metagenomic profiling of intestinal microbiome in extreme ageing. Aging *5*, 902-912.
91. Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Federhen, S., et al. (2011). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res *39*, 38-51.
92. Xue, G.-P., Johnson, J.S., and Dalrymple, B.P. (1999). High osmolarity improves the electro-transformation efficiency of the gram-positive bacteria *Bacillus subtilis* and *Bacillus licheniformis*. Journal of Microbiological Methods *34*, 183-191.

5.6 SUPPLEMENTARY MATERIAL

5.6.1 'Paris' medium protocol for *B. subtilis* WB800N transformation

Inoculate 50ml MG1 medium (per 50 mL; 0.5ml casamino acids (2% (w/v)); L-tryptophan to 50 µg/mL) with a fresh colony of *B. subtilis* WB800N in a 250 mL Erlenmeyer flask and adjust the OD_{650} to ~ 0.7. Incubate the culture for 3 hrs at 37 °C with shaking at 200 rpm. Dilute the culture 10-fold using 50 mL pre-warmed MG2 medium (per 100 mL; 0.5 mL casamino acids and 10 µg/mL L-tryptophan) and incubate for a further 90 min. To freeze competent cells pellet the cells by centrifugtation at 5000 g for 5 min at 20 °C. Remove the supernatant and mix it with 8.5 % (v/v) glycerol and 0.5% (w/v) glucose. Immerse a micorcentrifuge tube in liquid nitrogen and resuspend the cell pellets in the supernatant-glycerol-glucose mixture. Aliquot the suspension into the microcentrifuge tubes and store at – 80 °C. To transform the cells, thaw an aliquot of cells quickly in a 37 °C water bath and dilute 10-fold in transformation buffer (2 mL （MM x5 media: per litre:$(NH_4)_2SO_4$, 10g; K_2HPO_4, 74g; KH_2PO_4, 27g; trisodium citrate) 0.2mL 1 M $MgCl_2$ 0.25 mL 20 % w/v glucose 0.1 mL 0.1 M EGTA and H_2O to 10 mL) with gentle mixing. Add 100 µL diluted cells to 1-10 µL plasmid (up to 1 µg) and incubate for 20 min at 37 °C with no shaking. Add 500 µL of LB medium and incubate for 1.5 hrs at 37 °C with shaking at 150 rpm and plate transformants on selective media.

5.6.2 Spizizens protocol for *B. subtilis* WB800N transformation

Inoculate *B. subtilis* WB800N, which has been growing for a few days, in 5 mL HS medium (10 mL 10 X spizizens salts: per Litre 20 g $(NH_4)_2SO_4$, 140 g K_2HPO_4, 60 g KH_2PO_4, 1 g sodium citrate, make up to 1 L with ddH_2O and autoclave and add 0.1 mL 1 M $MgSO_4$. 2.5 mL 20 % w/v glucose, 5 mL 0.1 % w/v L-tryptophane, 1 mL 2 % w/v casein hydrolysate,5 mL 10% w/v yeast extract, 10 mL 0.4 % histidine and 0.1 mL 2 % ferric ammonium citrate and make up to 100 mL with ddH_2O) and incubate at 37 °C

with shaking at 150 rpm for 16 hours. Inoculate fresh HS medium (50 mL) with 0. 5 mL of the overnight culture and shake at 37 °C at 200 rpm and monitor the culture using OD_{650} until the culture enters early stationary phase (usually 4-5 hours). After such time, inoculate 20 mL starvation medium (10 mL 10 x Spizizens salts, 2.5 mL 20 % w/v glucose, 0.5 mL 0.1 % w/v L-tryptophan, 0.5 mL 0.2 % casein hydrolysate, 5 mL 2 % w/v yeast extract, 0.25 mL 1 M $MgCl_2$ and 0.05 mL 1 M $CaCl_2$ and make up to 100 mL with ddH_2O and autoclave) with 1 mL cells and incubate at 30 °C with shaking at 100 rpm for 2 hrs. Add 10 µM 0.1 M EGTA and incubate for 5 min at room temperature. Add up to 1 µg plasmid and incubate for 2 hrs at 37 °C while shaking at 150 rpm. After incubation, plate transformants on selective media.

5.6.3 MD Medium protocol for *B. subtilis* WB800N transformation

Prepare MD medium with 10 X PC (per Litre: 107 g K_2HPO_4,60 g KH_2PO_4, 10g Na_3 citrate.$5H_2O$ and 8.5 g $Na_3.2H_2O$) and add 5 mL to MD medium (per 50 mL: 2.5 mL 40% w/v glucose, 1.25 mL 2 mg/mL L-tryptophan, 250 µL 2.2 mg/mL ferric ammonium citrate, 2.5 mL 50 mg/mL L-aspartate, 150 µL1M $MgSO_4$ make up to 50 mL with ddH_2O and pH to 7.0 with KOH and filter sterilise). Add MD medium (10 mL) to a sterile 250 mL flask and supplement with 50 µL 20% (w/v) casamino acids. Store the remaining MD medium at 37 °C. Inoculate the flask of MD with a small loopful of freshly streaked *B. subtilis* WB800N. Incubate the culture at 37 °C with vigorous shaking until the OD_{600} reaches 1-1.5. To induce competence, add MD medium (10 mL) lacking casamino acids and shake at 37 °C for a further 1 hr. Add competent cells (800 µL) to up to 1 µg/mL plasmid and shake at 37 °C for 20 min and add 25 µL casamino acids (20% w/v). Shake at 37 °C for 1 hr and plate on selective media.

5.6.4 Electrotransformation protocol of *B. subtilis* WB800N

Electrotransfomation of *B. subtilis* WB800N was attempted using the protocol as described by Xue et al., [92]. Briefly, an overnight culture of *B. subtilis* WB800N grown in LB medium supplement with 0.5 M sorbitol was diluted 16-fold into fresh LB-sorbitol medium and incubated at 37 °C with shaking at 250 rpm until the culture reached an optical density of 0.85-0.95 (OD_{600}). Cells were incubated on ice for 10 min and centrifuged at 4,000 g for 5 min at 4 °C and washed four times in ice-cold

electroprotation medium (0.5 M sorbitol, 0.5 M Mannitol, 10 % (v/v) glycerol). After washing, cells were suspended in 1/40 of the starting culture volume with cold electroporation medium. Plasmid DNA (>50 ng was added to 60 µL of cells in a cooled electroporation cuvette (0.2 cm-gap, Bio-rad Laboratories, Ltd, Herts, UK and kept on ice for 1 min. An electrical pulse was applied to the cells using a Bio-rad Gene Pulser electroporation unit (Bio-rad) at 1.6 kV, 200 Ω resistance and 25 µF capacitance. Recovery medium (LB with 0.5 M sortbitol, 0.38 M Mannitol) was added to the cells after electroporation and the mixture was transferred to 15 mL falcon tubes (VWR) and incubated at 37 °C with shaking at 200 rpm for 3 hours before plating on selective media and overnight incubation.

5.6.5 Plasmid (pNZ44) digest

pNZ44 was routinely digested with *Nco*I or *Sph*I restriction enzymes. To ensure plasmid had been efficiently digested, product was run on a 0.8 % w/v agarose gel alongside a 2-log DNA ladder (NEB). Linearised plasmid should have been ~ 3 kb in size.

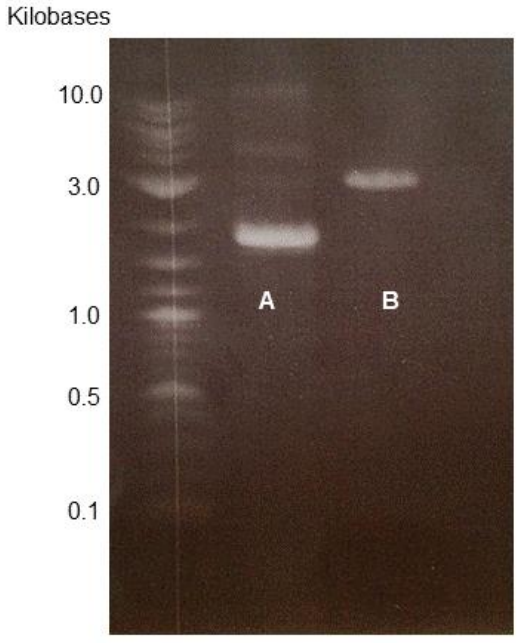


Figure S 5.1 A; non-digested (circular) pNZ44, B; plasmid digested with NcoI

5.6.6 Genomic DNA partial digest

After DNA extraction from faecal samples, gDNA was partially digested with *Fat*I or *NIa*III. Fragments of 3-10kb were cut out of the gel and purified.

Figure S 5.2 Partial digest of gDNA using *Fat*I 4 cutter restriction enzyme. DNA fragments of size 3-10kb were excised from the gel using a sterile scalpel and prepared for succeeding FM library construction steps.

6.0 THE EFFECT OF STORAGE CONDITIONS ON THE REPRODUCIBILTIY OF PROTEIN CONCENTRATION AND PROTEASES FOR ANALYSIS OF THE FAECAL MICROBIOTA

6.1 INTRODUCTION

The human gut harbours microorganisms whose collective genomes out numbers our own by at least 100 to 1 genes [1]. The implementation of 'omic' based technologies, in particular, metagenomics have hugely advanced our understanding of the human gut, and have allowed the elucidation of the core, broad adaptive functions of the gut microbiota For example, Qin and colleagues defined the minimal gut metagenome and unearthed core species that were indicative of a healthy gut [2] Gloux and colleagues implemented a functional metagenomic approach to study β-glucuronidase activity in the human gut microbiota and revealed a class of β-glucuronidase enzymes that were specific to species of the gut suggesting these enzymes were part of a functional core and had adapted specifically for life in the human gut [3]. Similarly, Jones and colleagues implemented a functional metagenomic approach to screen the human gut metagenome for bile salt hydrolases (BSH). Functional BSH were identified in all major divisions of bacteria and it was deduced that BSH activity appeared to be a conserved adaptation to life in the human gut [4]. However the implications of such functions in specific bacteria and their specific interaction with host cells in homeostasis or in the development of disease are far from being understood. Consequently a coupling of a variety of experimental approaches is required to elucidate these functions and interactions, including functional metagenomics, deep sequencing approaches and importantly function-based assay. Studying the specific functions of bacterial molecules and enzymes, behaving as they would in their normal environment, may give us a much better understanding of what the genes identified using the powerful deep-sequencing based techniques, are actually capable of conducting. Pathogens have been identified and dealt with by understanding and isolating the molecules and their mechanisms of their interaction with our cells that consequently contribute to pathogenesis. By this strategy we have tried to develop novel ways to combat the pathogen. Therefore perhaps it is fair to say that a more directed function based analysis of microbial functions (i.e. not just sequence-based analysis alone) would be a key step in unravelling the role of the gut microbiota in health and disease [5].

Faecal samples are often used as a surrogate for the gut microbiota [6] and offer a non-invasive, cost-effective and repeatable means of accessing the gut microbiota from both humans and animals. The use of faecal samples in pinnacle projects such as that conducted by Eckburg *et al.,* [6] on the human large intestine have provided a wealth of information regarding the gut microbiota. Information generated from Eckburg's study has shown that mucosa-associated bacterial communities, isolated from different positions along the human gut, and faecal samples, do differ from each other within individuals. UniFrac analysis revealed a distinct clustering of the data between mucosal and faecal samples within an individual compared to between different individuals. The insights gained from this study clearly indicate that faecal samples can be used as an easy, safe and reputable proxy for describing both intra- and inter-personal differences in the gut microbiota composition and function [7].

Arguably one of the most important steps in metagenomic studies is that of the isolation of DNA to obtain a non-bias representation of the microbiota in that particular sample. Optimal methods for DNA extractions to confer the best yield, high quality, large fragment size and unbiased DNA that is representative of the entire microbiota [8] have been studied in depth and these processed have been discussed at length in Chapter 5.0 (and for more information please see [8-12]). Long term storage of DNA extracted from faecal samples as well as other environments have also been extensively studied [13-15]. For example Lauber *et al*., used barcoded pyrosequencing to assess the stability of the gut microbial community using 16S rRNA gene analysis from faecal samples stored at 20°C, 4°C -20°C and -80°C for 3 and 14 days. They concluded that microbial communities were not significantly affected by storage temperature or duration [16] Carrol *et al.,* took this further by conducting a comprehensive analysis of the stability of the human gut microbiota using 454 pyrosequencing from DNA sample stored at room temperature for 24 hours and after storage at -80 °C for 6 months having also determined the microbial community structure as immediate as possible to DNA extraction to determine whether there could be an immediate loss in certain faecal bacteria before the sample was processed. It was however, concluded that samples did not significantly change after storage in both conditions suggesting that the microbiota was retained and there was no significant loss in diversity [17]. An abundance of

research has been conducted in this area, and it is generally considered that DNA sample and other enzymes are best kept at sub-zero temperatures.

As the frequency of metaproteomic studies increases, as does the necessity to develop optimal protein extraction methods [18-20]. However, proteins can exist in many different biological forms, even if isolated from the same environment and so developing a general extraction method for proteins is particularly difficult [21]. Additionally, protein samples for metaproteomics often do not need to maintain their function, whereas clearly in activity-bases studies, this is of paramount importance.

The use of faecal samples, from both human and animal subjects, to obtain DNA and enzymatic functions has been proven to be an extremely useful and invaluable tool for studying the functions of the gut microbiota. The direct study of microbial functions such as specific enzyme activities as they behave in their natural environment may also be of significant use in furthering our understanding of microbe-host interactions. Studies on the functional activity or levels of certain cells, protein/enzymes, hormones etc. has provided invaluable insights into the broad areas of gut health, disease aetiology and disease pathogenesis. For example, there are a number of measurable faecal biomarkers for intestinal inflammation such as calprotectin [22] and lipocalin 2 [23]. Faecal hormones (such as glucocorticoids) levels can be used to monitor stress in animals [24]. Monitoring faecal microbial enzymes activities such as that of beta-glucuronidase and beta-glucosidase have helped to show how diet can change physiological activities of the gut. Freeman and colleagues showed that dietary fibre decreased faecal beta-glucosidase activity which appeared to have a protective role against the development of experimental colitis in a rat model of colon cancer [25]. Similarly, Flores et al., observed bacterial taxonomic correlations with decreased or increased levels of beta-glururonidase and beta-glucosidase demonstrating how alterations in the gut microbiota also alters their overall functional capacity and this has implications on host gut health.

A number of studies have focused on carbohydrate metabolising functions of the gut microbiota, however, the gut harbours an expanse of other molecules both endogenous and dietary. The degradation of such molecules by the host ,and by the gut microbiota,

particularly the consequences of degradation regarding host health, remain relatively unexplored. Macfarlane and colleagues hypothesised that as large quantities of soluble protein, ammonia and volatile fatty acids had been found in the large intestine, that proteolysis must be occurring at substantial levels in the gut [18]. Their culture dependent study and a culture dependent study conducted previously as part of this thesis (please see Chapter 5.0) on the bacterial contribution to proteolysis in the large intestine identified, several proteolytic species of bacteria isolated form the human gut, including *Bacteroides vulgatus, B. fragilis, B thetaiotaomicron*, spp. *Propionibacterium* spp. *Streptococcus* spp and *Staphylococcus* spp. as well as *Enterococcus faecalis* and *Bifidobacterium* spp. demonstrating the bacteria are likely to be significantly contributing to proteolysis in the gut. A comparison of faecal protease activity before and after bacterial cell disruption has also shown an increase in protease activity indicating large quantities of proteases are released by bacteria in the gut [26].

Digestion of dietary protein leads to an array of end products including short chain fatty acids, amines, phenols, indoles and thiols; some of these compounds have been shown to have toxic properties [27] and have been implicated in diseases of the gut including Irritable Bowel Syndrome (IBS) and Inflammatory Bowel Disease (IBD) [28]. Protein degradation by microorganisms in the human gut has also been linked with colon cancer [29] and specific bacterial proteases have been shown to influence invasiveness of colon tumour cells [30] and may be directly linked to progressing the cancer [31].

Microbial proteases represent approximately 5% of the genomes of infectious organisms [32] and approximately 2% of the genomes of commensal [33] and their contribution to disorders of the human gut are discussed at length in chapter 7.0.Proteolytic activities have been studied by direct methods extensively by labs such as Macfarlane and colleagues and have given us valuable insights into functionality of the healthy gut [18] and to some extent, on the unhealthy gut [28]. Most studies on proteases and how they may be impacting the host are carried out on pure cultures of specifically chosen microorganisms which may not actually be numerically dominant in the gut [34-36]. While these studies are incredibly insightful, and demonstrate how significant microbial proteases and their interactions with host cells are, it is also useful to study enzymatic functions as a separate entity i.e. proteases, or as a collective; the degradome, as they

are found in a particular environment at any one time point. Stool samples offer this non-invasive insight into proteolytic activity of the human gut microbiota as it may be occurring *in situ*.

Some of the major aims of this thesis was to conduct a comprehensive comparison of faecal protease activity in a cohort of IBD patients compared to healthy volunteers with the hypothesis that proteases may be acting as virulence factors in IBD, or protein degradation by the gut microbiota in IBD patients maybe indirectly contributing to gut inflammation through a variety of pathways and different substrate metabolism (Chapter 7.0). However, to give a comprehensive and non-bias insight, all the analysis cannot be carried out after fresh extractions from fresh stool samples. Therefore, prior to embarking on a study of specific enzyme activity and it's clinical significance certain parameters have to be optimised to minimise variability in results [5]. As previously mentioned a substantial amount of work has been conducted on DNA and increasing amount of protein for proteomics. However, optimal methods for extraction of enzymes from faecal samples, and their reproducibility after long term storage remains relatively ignored. Well established protein storage methods often involve the irreversible inhibition or denaturation of proteases to prevent proteolytic degradation of protein that do not require a functionally active protein. Studies are often focused on obtaining the best yield and purity and while purity and yield are important here, the protein still needs to be intact and fully functioning, hopefully as they would be *in situ*.

The parameters that may give variable results first need to be identified and controlled to ensure that results are reputable. It is widely known that maintaining protein for extended periods of time can cause stability issues. Protein becomes susceptible to denaturation and alterations upon extraction; there are many effects that can cause a protein to become denatured and thus cause a protein extract to decrease dramatically in concentration. Such factors include dramatic changes in temperature, repeat freeze thaw cycles, contamination and the activity of proteases. Temperature changes can cause denaturation by disrupting the secondary and tertiary structure of proteins, while disruption of protein structure is normally considered a result of temperature increase whereby hydrogen bonds and hydrophobic interactions are disrupted, which is why it is more useful to maintain a protein sample, purified or mixed at cooler temperatures.

Contamination is also an issue, therefore normally some form of antimicrobial agent such as sodium azide (NaN_3) is added, but care must be taken that this does not affect the protein's activity. For long term storage of protein extracts some sort of cryoprotectant such as glycerol has to be added at concentrations of 10-50% (v/v) and samples should be frozen and thawed as quickly as possible to avoid denaturation of protein. Reducing agents such as Dithiothreitol (DTT) may also be added to prevent oxidation of cysteine as the thiol group of cysteine is highly reactive.

In this study, buffers where designed so as not to interfere with protein concentration measurements (purified bovine serum albumin ,BSA, is often added to protein samples for increased stability and to prevent binding to the storage vessel, however in this instance, BSA would then alter protein concentration estimations so could not be added). Protease activity was assessed following long term storage of a crude protein extract from human faecal samples. Alterations in protein concentration were also monitored as well as the effect of repeated freeze-thaw cycles was determined with an aim to specifically study faecal protease activity levels following various lengths of storage. The effect of the long term storage of whole faecal specimens as a means of accessing the faecal degradome after long periods of storage was also determined.

Unfortunately, the principles that must be applied for storage of basic protein extracts cannot be adhered to in a study where the sole purpose is to assess protease activity. For example, many standard protein storage practices involve addition of a protease inhibitor cocktail to prevent degradation of protein. Obviously, a protease inhibitor cannot be added to a sample in which the proteases themselves are of sole interest, so it will have to be taken into account that there may be some degradation of sample protein by endogenous proteases. Also, any compounds that may affect protein concentration measurements (e.g. BSA is often added to stabilise protein solutions) could not be used as this would introduce biases in subsequent assays.

AIMS

Previous research documented here (Chapter 5.0) has focused on the use of DNA and metagenomics techniques to study the proteolytic potential of the gut microbiota however it has become apparent that functional analysis is also necessary. The overall aim was to determine what happens to protease activity when faecal samples are stored to enable effective and consistent studies of the gut microbiota degradome (Chapter 7.0) and this was achieved by a series of sub-aims which were as follows:

1. Determine how bead beating affects protein recovery from faecal samples.

2. Determine whether the preservative, sodium azide (NaN_3), has any impact on protein extraction or protease activity.

3. Determine how PBS with a variety of additives such as glycerol and DTT affects protein concentration and protease enzyme extracted from a faecal samples during long-term storage.

4. Assess whether storing whole faecal samples at sub-zero temperatures allows reproducible measurements of protein concentration and protease activity for use in future experiments.

6.2 RESULTS

6.2.1 Effect of number of bead beatings on protease activity

The effect of the number of bead beating steps on protease activity are shown in figure 6.2.1 Results represent the mean ± S.E.M of triplicate results from participants (n = 3). Protein concentration became significantly different after 3, 4, 5 and 6 bead beating steps (figure 6.2.1, $P < 0.05$) when compared to activity after one beating. Protease activity was also significantly higher after 3 beatings when compared to 2 beatings however no difference was observed after 4, 5 and 6 beatings compared to 2 beatings suggesting that more than 3 beatings may be deleterious to protease activity.

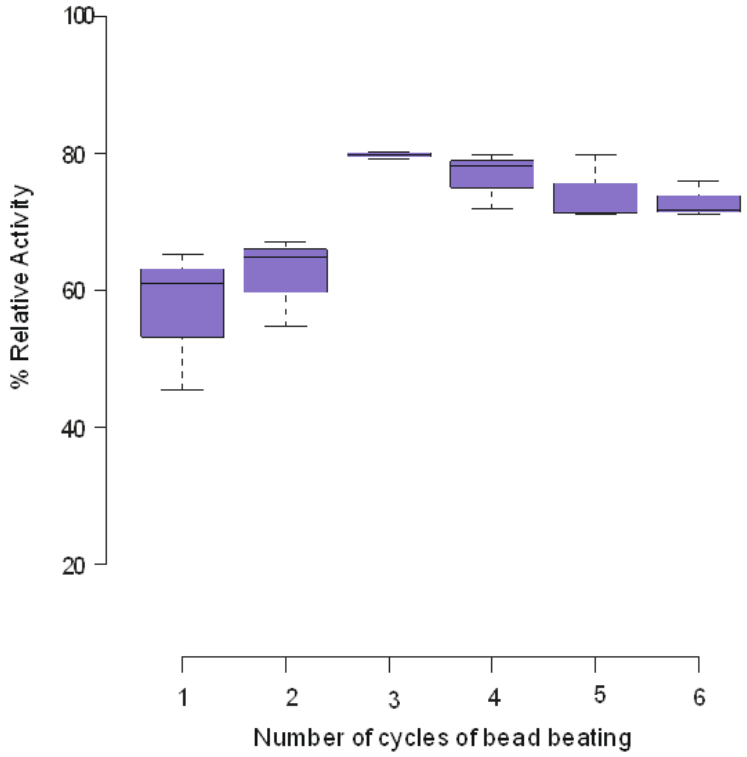


Figure 6.2.1 A boxplot of the mean of triplicate data + standard error of the means. Statistical significance was determined by non-parametric multiple comparison testing using a Holm adjustment method to control for family wise error rate. Significantly higher protease activity was observed after 3, 4, 5 and 6 rounds of bead beating (P = 0.0065, 0.0196, 0.0430 and 0.0639 respectively) when compared to 1 round of bead beating. Significantly higher activity was also observed after 3 beats compare to 2 beats (P= 0.0377).

6.2.2 Effect of NaN_3 on protease activity

The effect of sodium azide on protein concentration after extraction and protease activity was determined by conducting an azo-casein hydrolysis assay from a normalised input of protein (1 mg/mL). To remove bias from differences in protease activity between individuals, % relative activity was calculated separately for each sample based on 100% activity being the hydrolysis of azo-casein after double the incubation time (4 hours). Data was found to be normally distributed and statistical significance was determined by the student's t- test in R software.

No significant difference found between protease activity (figure 6.2.2) when NaN_3 was used and when it was absent. Suggesting it does not affect these measurements at this concentration and so can be used in a buffer to store proteases as an antimicrobial agent.

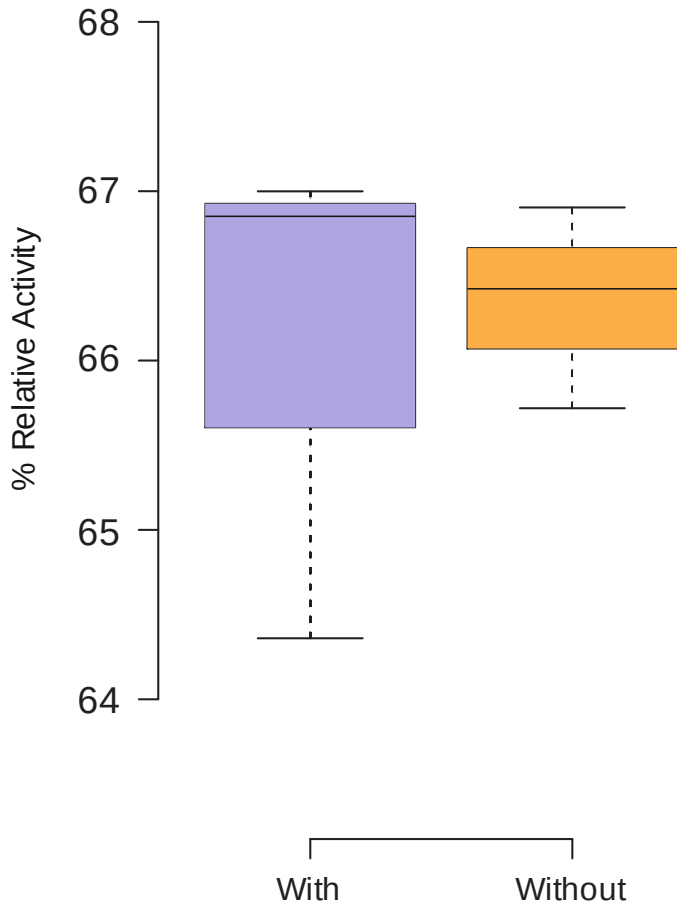


Figure 6.2.2 Effect of 0.05% NaN_3 on protease activity. Centre lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles. Statistical significance was determined by the student's t-test. ($P>0.05$)

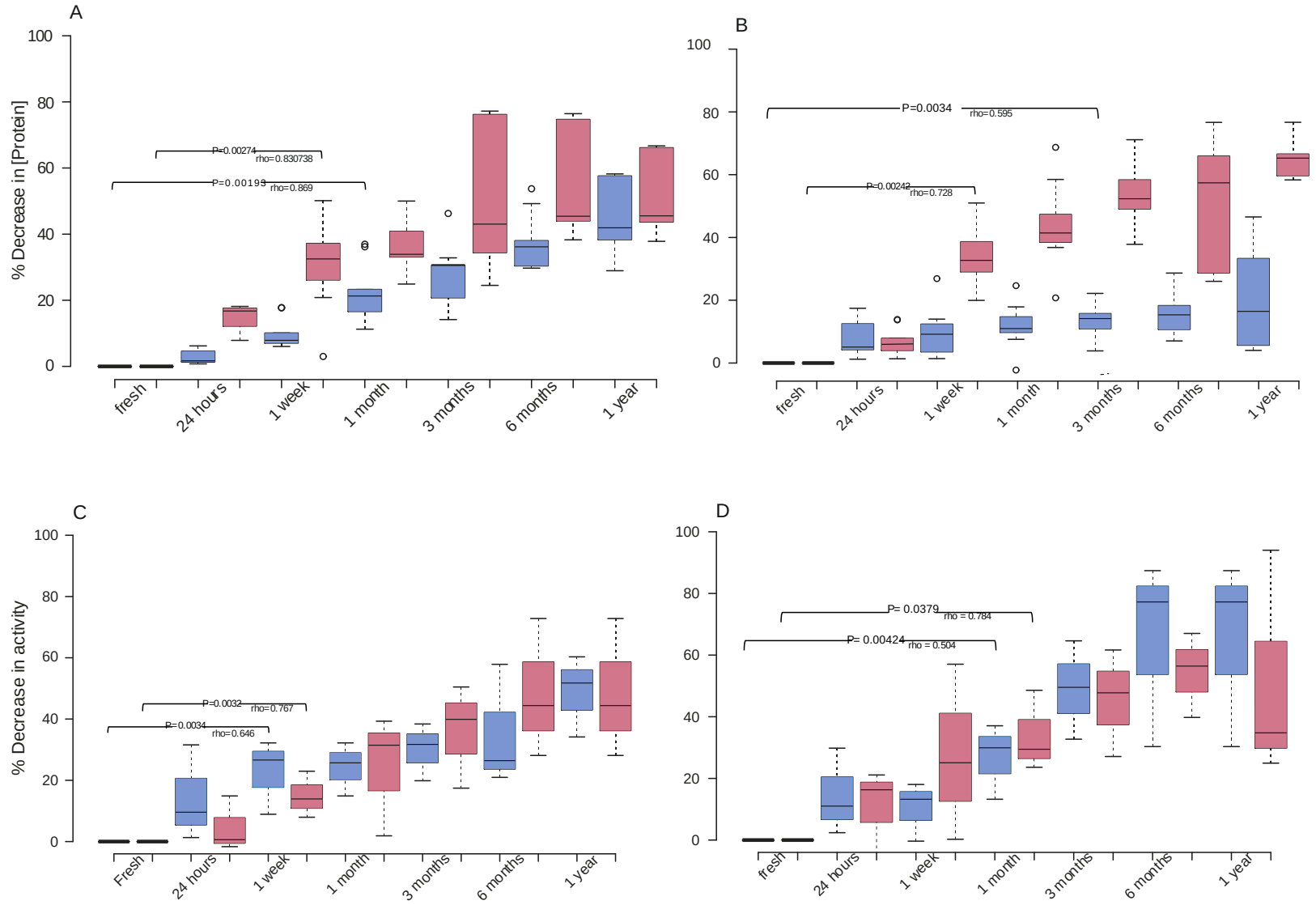
6.2.3 Stability of protein concentration and protease activity of faecal protein extract over time

To monitor the reproducibility of protein concentration estimates after extraction and storage over time, a crude protein extract was prepared from fresh faecal samples from 3 individuals as described in methods. Protein concentration measurements were recorded by Bicinchoninic acid assay (BCA) assay after fresh extraction, 24 hours of storage in each buffer and following measurements at 1 week, 1 month, 6 months and finally 1 year. As protein concentrations are diverse according to different individuals, in order to be able to compare the data as a cohort, the % decrease in protein concentration was calculated for each samples compared to the protein concentration measurement taken after the initial extraction. The resultant data was used to determine whether there was a significant correlation between the level of % change in protein concentration and the duration of storage and the original data can be observed in supplementary materials (figure S1). Firstly, non-parametric correlation analysis was implemented to determine whether there was a statistically significant correlation and non-parametric pairwise comparisons using the Mann Whitney –U test were conducted in R to elucidate the impact of length of storage and buffer type on protein concentration estimates. All samples exhibited a significant correlation between decrease in protein concentration and storage time, regardless of buffer type displayed by the results of the spearman rank correlation coefficient (rho)(figure 6.2.3 A and B) which was conducted to determine if there was correlation between the two variables (length of storage and protein concentration or protease activity). Both the extracellular and total protein extracts stored in buffer 1 displayed a significant % decrease in protein concentration after 1 week of storage at -20 °C ($P =<0.003$ for both). Total protein extract stored in buffer 2 showed a significant % decrease in protein after 1 month ($P<0.002$) and extracellular protein stored in buffer 2 showed a significant % change in protein concentration after 3 months ($P < 0.004$).

To determine the stability of faecal proteases isolated after crude protein extraction over time following extraction in, and subsequent storage in two different buffer compositions, protease activity measurements were conducted using a colorimetric azo-casein substrate with 100 µl of protein sample which had been diluted to 1 mg/mL. It was observed that different individual's display various levels of protease activity and so

in order to be able to compare the data as a cohort, the % change in protease activity over time was calculated for each samples compared to the protein concentration measurement taken after initial extraction. The resultant data was used to determine whether there was a significant correlation between the level of % change in protease activity and the duration of storage. Firstly non-parametric correlation analysis was implemented to determine whether there was a statistically significant correlation between length of storage and protease activity and then non-parametric pairwise comparisons were conducted in R to elucidate the impact of length of storage on protease activity and also whether the different buffers showed differences in their ability to maintain protease activity.

All samples demonstrated a significant correlation between length of storage and the % change in protease activity as demonstrated by the spearman rank correlation coefficient calculated (figure 6.2.3 C and D) suggesting that protease activity in all samples was affected deleteriously by the length of storage. Total protein stored in buffer 1 and buffer 2 displayed a significant decrease in protease activity after 1 week of storage ($P = 0.0032$ and $P = 0.0034$ respectively) whereas the extracellular protein protease activity displayed more stable protease activity which only began to show significant changes after 1 month of storage in buffer 1 ($P = 0.00424$) and after 3 months of storage when stored in buffer 2 ($P = 0.00379$). These data indicate that both buffers are only appropriate for storing total protein isolated from faecal samples for a short period of time (1 week) before they begin to show changes from their original activity and therefore further function-based analyses on these samples would not be reputable. Extracellular proteases appear to be more stable in these buffers.


A
% Decrease in [Protein]
100
80
60
40
20
0
P=0.00274
rho=0.830738
P=0.00193
rho=0.869
fresh
24 hours
1 week
1 month
3 months
6 months
1 year
B
P=0.0034
rho=0.595
P=0.00242
rho=0.728
fresh
24 hours
1 week
1 month
3 months
6 months
1 year
C
% Decrease in activity
P=0.0032
rho=0.767
P=0.0034
rho=0.646
Fresh
24 hours
1 week
1 month
3 months
6 months
1 year
D
P= 0.0379
rho = 0.784
P= 0.00424
rho = 0.504
fresh
24 hours
1 week
1 month
3 months
6 months
1 year

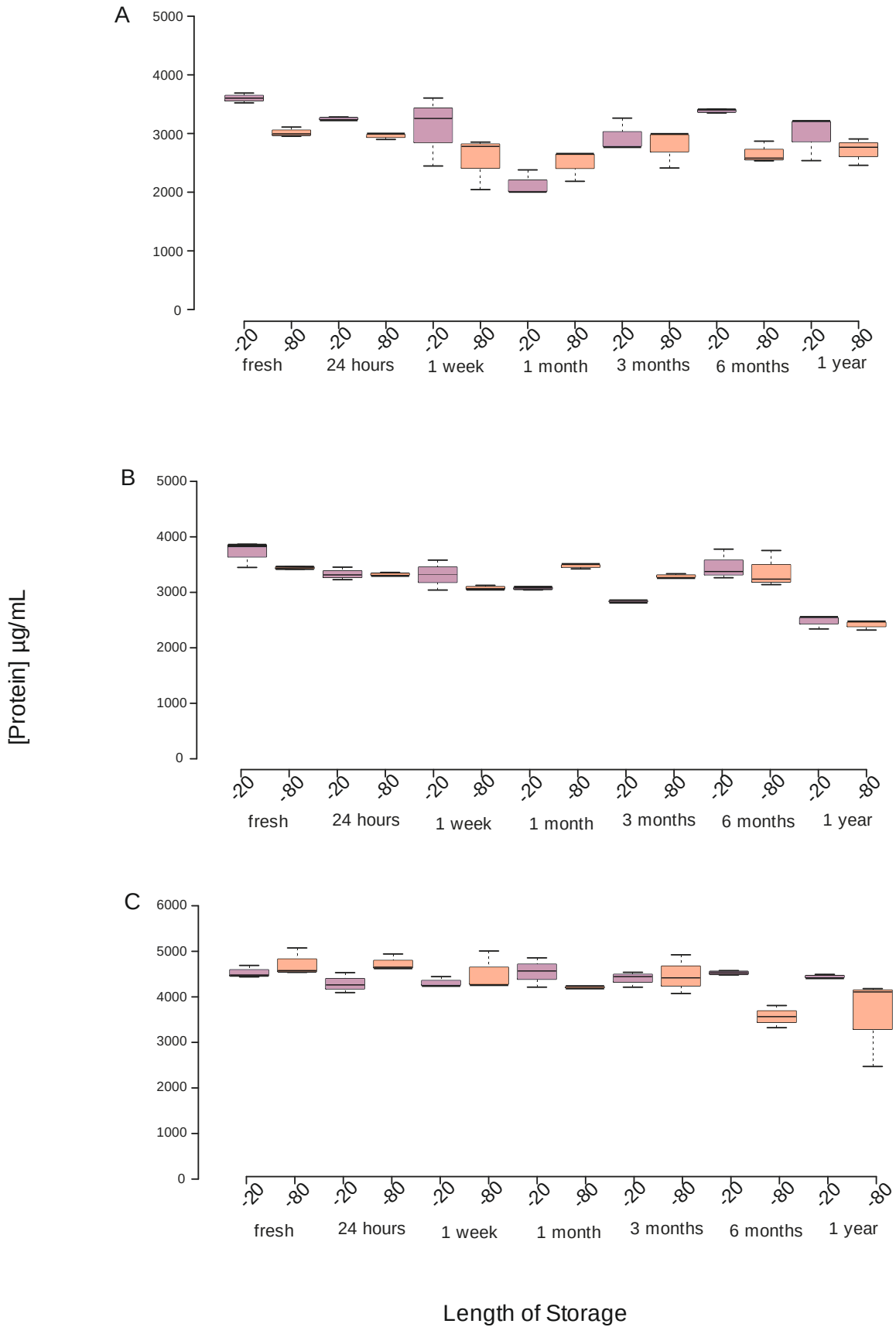
1

←from previous page

Figure 6.2.3. A= the relationship between length of storage and the % decrease in protein concentrations of crude protein extract after bead beating (total protein) in two different buffers; blue=buffer 2 and pink = buffer 1. **B** = decrease in extracellular protein after storage. **C**=The relationship between length of storage and the % decrease in protease activity in total protein samples stored in the two different buffer and normalised to 1 mg/mL. **D**= Extracellular protein protease activity. Centre lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by dots. n = 9 sample points.

6.2.4 The effect of freezing temperature of whole faecal samples and reproducibilty of protein concentration and protease acitivity

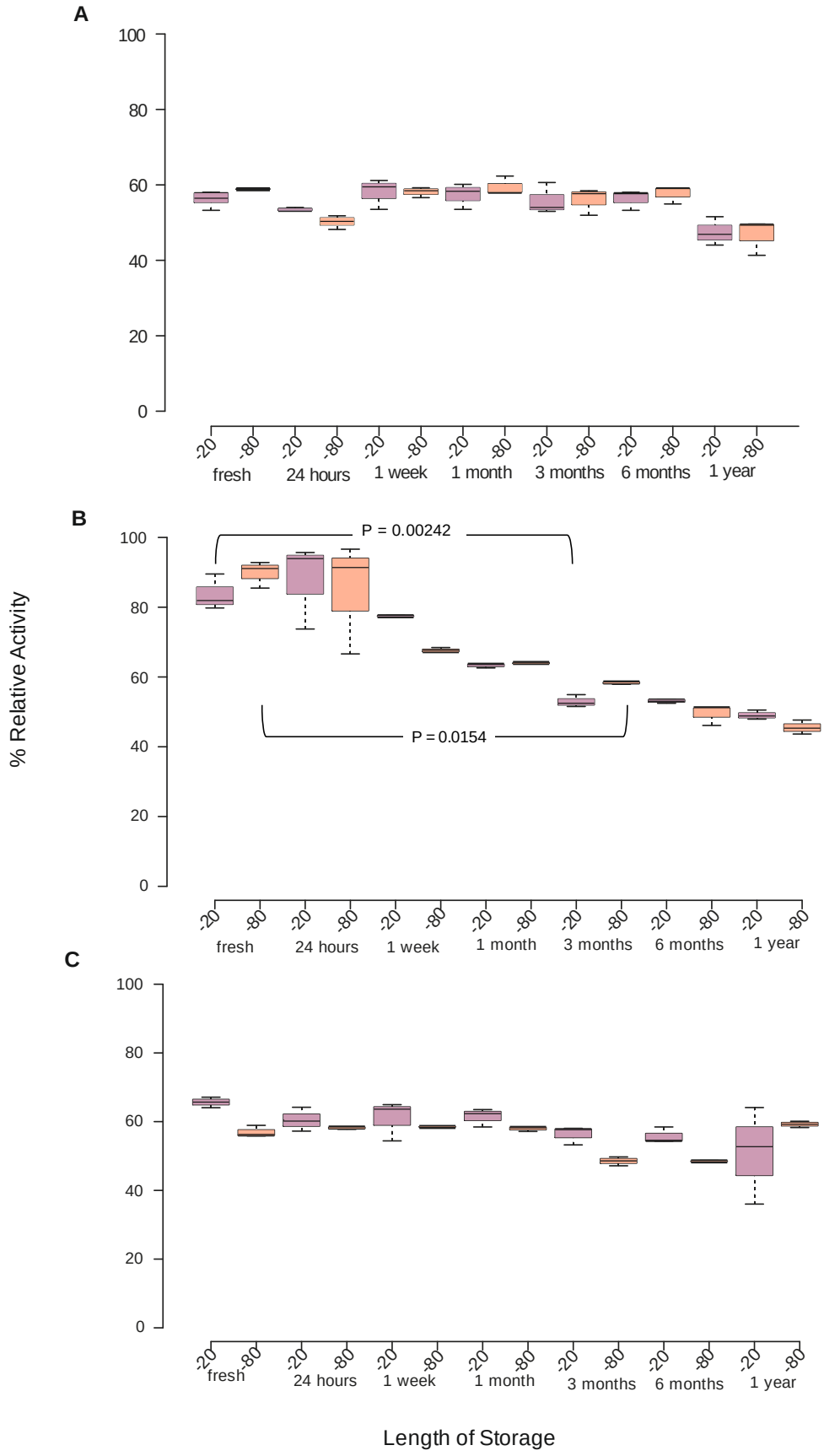
Protein concentration was not affected by length of storage for any of the samples collected from the three different individuals (figure 6.2.4) nor was it affected by storage temperature suggesting that faecal protein remains stable when frozen as part of the original biologcal sample at sub-zero temperatures. 2/3 samples displayed no signifciant alterations in protease acitivty following normalisation to 1 mg/mL protein over the entire period of the study at both of the storage temperature (figure 6.2.5, A and C), however samples collected from individual 2 exhibited a significant change in proteolytic activity follwing three months of storage at both sub-zero temperatures (figure 6.2.5, B).


A
B
C
5000
4000
3000
2000
1000
0
6000
-20
-80
fresh
24 hours
1 week
1 month
3 months
6 months
1 year
[Protein] µg/mL
Length of Storage

←*from previous page*

Figure 6.2.4 Effect of freezing temperature of whole faecal sample protein concentration estimates

The effect of freezing temperature and the length of storage on whole faecal samples and their protein concentration. Boxplots indicate the results of triplicate data, Centre lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles. The data were analysed by Wilcoxon signed ranked tests in R software to assess whether protein concentration deviated from the concentration measured upon fresh extraction.


A
B
C
100
80
60
40
20
0
% Relative Activity
P = 0.00242
P = 0.0154
-20
-80
fresh
24 hours
1 week
1 month
3 months
6 months
1 year
Length of Storage

←*from previous page*

Figure 6.2.5 Relative protease activity (%) after protein extraction of faecal samples stored for one year at -20 °C and -80°C. The effect of both freezing temperature and the duration of storage of faecal samples on protease activity after normalisation to 1 mg/mL protein following protein estimations. Boxplots indicate the results of triplicate data, centre lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles. % relative activity was determined using 1 mg/ml proteinase k incubated with the same substrate for 2 hours. The data were analysed by Wilcoxon signed ranked tests in R software to assess whether protease activity deviated from the activity measured upon fresh extraction. The time at which activates became significantly different to original activity (fresh samples) are indicated along with their respective P-values.

6.3 DISCUSSION

The Gut microbiota has an enormous diversity in terms of both phylogeny and functionality [6, 37], therefore a diverse array of techniques including'-omic' approaches as well as function-based assays need to be implemented if we are to fully understand the impact they are having on us, as their hosts. However, as highlighted by work such as that by Flores and colleagues in their study of the reproducibility of β-glucuronidase and β-glucosidase enzymes in faecal samples after extraction and storage [5, 38], researchers cannot begin a study without first assessing the parameters that may be affecting the gut microbiota and the molecules they produce that may be under scrutiny in a particular study. Results should also be reputable and a normalisation of the way certain samples are stored for particular studies is required to reduce bias and to allow for comparison of studies across different research laboratories.

An ultimate aim of this research is to assess proteolysis in the human gut my implementing novel tools to comprehensively analyse faecal protein samples for general protease activity and specific protease activity (Chapter 7). However, it is impossible to carry out these comprehensive analyses immediately on fresh protein extract. Therefore, before embarking on such research, certain parameters have to be met to try to minimise bias and to ensure future results are robust. In this chapter, the aim was to assess the stability of protein extracts isolated from faecal samples over time and to deduce which buffer composition, and whether freezing faecal specimens before processing, was appropriate for maintaining protein stability with an overall objective of maintaining the stability of the proteases within the sample since there are no ways of isolating unknown proteases from complex samples, such as human faeces as of yet. Due to the aims of this study, this meant that there could be no component of the buffer that would affect protein measurements nor inhibit or enhance potential protease activity. Therefore, the composition of the buffer could not include carrier proteins such as BSA which normally ameliorates the problem of low-level binding to storage vessels, metal chelators such as EDTA as this could affect the activity of some proteases, protease inhibitors because firstly they themselves are proteins and secondly, they will inhibit the activity of the proteases present, the prominent aspect of this study. It was also found that the reducing agent dithiothreitol (DTT), which prevents oxidation of cysteines,

severely skewed the protein concentration measurements so could not be used for further experimentation. Consequently, it was decided that the final buffer compositions to compare would be simply PBS supplemented with 0.05% (w/v) NaN_3 (buffer 1 compared to a 20% glycerol-PBS buffer also supplemented with 0.05% (w/v) NaN_3 (buffer 2). NaN_3 is a potent antimicrobial and so it's inclusion in the buffer was essential for activity as a broad spectrum anti-microbial to prevent microbial contamination of faecal protein extract. It was shown that NaN_3 did not interfere with protease activity of the samples (Figure 6.2.2) and was successful in preventing microbial contamination evidenced by plating faecal protein extracts on to nutrient rich media and no bacterial growth occurring.Therefore it was concluded that NaN_3 was a useful component of a buffer for storage and maintenance of protease enzymes extracts in a complex biological sample.

The number of bead beating steps for optimal protein extraction was deduced by bead beating a sample from 1 to 6 times. The protein concentration and protease activity were measured for each of the bead beating samples and it was deduced that 3 bead beating steps was adequate to get maximum protein out of a sample without the addition of lysis enabling buffers (figure 6.2.1).

Buffer 1 and buffer 2 were compared over a period of 1 year with protein concentration and protease activity recorded immediately upon extraction, after 24 hours stored at -20°C and then after 1 week, 1 month, 3 months 6 months and finally 1 year to determine their expediency as buffers for protease storage. Following protein concentration measurements, protein concentrations were normalised to the same concentration for protease activity measurement. All samples in both buffers exhibited a negative correlation between both protein concentration and protease activity over time (figure 6.2.2) and no sample displayed repeatable protein levels or protease activity after 1 year nor 6 months of storage suggesting it would not be advisable to use either buffer for long term storage of protein extract for protease based studies. Assessment of shorter term storage revealed that after 1 week of storage total protein extract began to significantly decrease in concentration in both buffers, but extracellular protein remained more stable for 1 month in buffer 1 and 3 months in buffer 2 suggesting that buffer 2 was more suitable for longer term storage of extracellular protein but also that extracellular proteins are more stable than intracellular proteins perhaps because

proteins that are secreted from cells are adapted to the harsher, less controlled environment of the gut lumen.

The observed decrease in protein concentration is likely due to repeated freeze-thaw cycles, low-level binding to the storage tube and probably proteolytic degradation by proteases within the sample. The bicinchoninic acid solution for the assay contains a water soluble compound containing; bicinchoninc acid, sodium carbonate, sodium bicarbonate, sodium tartate and cupric sulphate pentahydrate and works in an alkaline environment. Peptide bonds in protein in the sample reduce ions (Cu^{2+}) in the cupric sulphate to Cu^{+} which form a purple complex with the Bicinchoninic acid salt [39]. When the protein samples increase in temperatures it is likely that the proteases present will also be contributing to the decrease in protein concentration by hydrolysis of the peptide bond, thus destroying protein and fewer peptide bonds, the fewer reduce cupric ions and so a decrease in in intensity of the purple compound as fewer cupric ions are available to bind the bicinchoninic acid salt. Protease activity however, was also diminished in all samples over the period of a year, in fact, protease activity from the total protein extracted began to deteriorate after just 1 week of storage. The most likely cause of this drop in activity is the loss of their native structure due to denaturation due to the freeze-thaw cycles and sub-zero temperatures can also affect protein structure [40]. Therefore, despite the presence of a cryoprotectant, the enzymes can still be rendered inactive. The extracellular protein maintained activity for up to 1 month when stored in buffer 1 and for 3 months when stored in buffer 2. Thus the glycerol is clearly playing a part in maintaining extracellular protease activity, but this also indicates that the extracellular protein is more robust than intracellular protein. It is likely that an increased concentration of glycerol would be more effective in maintaining both protein and consequently protease activity of faecal samples however; this would interfere with other assays (as well as the BCA assay and other protein quantification assays) and so is not ideal for studies aiming to evaluate protease activity.

A lot of studies involving the human gut microbiota involve the storage of whole faecal samples and different storage conditions can sometimes affect experimental outcomes. The effect of storage condition on faecal samples and concentration of other microbial molecules as a result of freezing have been studied [5, 41], but the effect of freezing entire faecal samples has on protease activity is yet to be deduced, until now. Here,

faecal samples from 3 healthy individuals were stored at -20 and -80°C. Protein was extracted and protein concentrations and general protease activity were measured immediately prior to any freezing, after freezing for 24 hours and subsequently after 1 week, 1 month, 3 months 6 months and 1 year. It was found that for all samples isolated from the 3 individuals there was no significant reduction in protein concentration even after a year of storage at both -20 and -80 °C (figure 6.2.3). Protease activity remained stable for the entire study period as well when faecal samples were stored at either temperature (figure 6.2.5) for two of the individuals studied. One sample however, displayed a reduction in activity after 3 months at both storage temperatures (figure 6.2.4 B) although it is noteworthy that this sample displayed higher protease activity than the other two samples suggesting self-proteolysis was playing a part in this reduction of protease activity.

Therefore, from this study it could be advisable that for future studies of protease activity in the human distal gut, if protein is extracted, a buffer comprising of a high concentration of cyroprotectant along with an antimicrobial, both of which do not alter protein concentration or protease activity, and storage at -20°C is suitable. If samples are required for extensive analysis which is likely to take more than 1 day, samples can be extracted in a glycerol-PBS buffer in a neutral pH and will remain stable for up to 1 week.

However, protein concentrations and protease activity measurements are more reliable when carried out analysis on frozen faecal samples. Following analysis of the results of this study it would be recommended that upon retrieval of faecal samples, 1g samples should be dispensed into appropriate, sterile storage vessels for future protein extractions. This approach will be more appropriate for storage in terms of space and importantly will help avoid repeat freeze thaw cycles of the entire sample and decrease the chance of contamination. Also, care must be taken with samples exhibiting significantly higher levels of protease activity compared to the rest of the cohort after storage for longer than 3 months as samples kept this long may not be representative of the gut microbiota from that individual. Though to conclude, human faecal samples offer a highly reproducible means of accessing protease activity of the human distal gut even after long periods of storage at sub-zero temperatures.

6.4 MATERIAL AND METHODS

6.4.1 Buffer composition and preparation.

All buffers were made up in molecular grade water (18 MΩ) and where possible steam sterlized at 121°C for 15 mins, where heat labile ingredients were presnt buffers were filter sterilised (0.2 micron). Four buffers were used in this chapter and the composition for each one is shown in Table 6.1. All reagents used were available from Sigma (Poole, UK) unless otherwise stated.

Table 6.1 Buffer Composition

Buffer	Buffer composition
1	PBS, 0.05% NaN_3
2	PBS, 10% glycerol, 0.05% NaN_3
3	PBS, dithiothreitol, 0.05% NaN_3
4	PBS, 10% Glycerol, Dithiothreitiol, 0.05% NaN_3

6.4.1 Collection of faecal samples

Assays to detect and quantify enzymatic activities were developed and optimized with faecal speicmens from laboratory volunteers (n = 3). Faecal samples were collected as descibed in the general materials and methods section (Chapter 2.0, section 2.3.1)

6.4.2 Protein extraction

Protein extractions were conducted as described in general methods section (Chapter 2.0, section 2.10.1).

6.4.3 Protein concentration measurements and Protease assays

These were conducted with the bicinchoninic acid assay and using the colorimetric azo-casein assay as described in the general methods section (Chapter 2.0, section 2.9.1).

6.4.4 Design, conduct and analysis of storage media for evaluation of protein yield and protease activity over time

Fresh material from one faecal sample was collected and divided into 8 subsamples (1 g). 1 sample for each buffer and each sample would be used for total protein extraction (i.e. with bead beating) and for extracellular protein only analysis (no bead beating) each containing 1 g faecal material. Each sample was allocated to a buffer (1,2,3 or 4) and this buffer was added to the faecal sample to prepare a 10% w/v faecal slurry which was conducted by mixing on a Vortex Genie 2™ until no clumps remained. To prepare the crude total protein extract the faecal slurry was divided into 2 mL RNAse and DNase free lysing matrix tubes (MP Biomedicals) containing 1.4 mm ceramic spheres, 0.1 mm silica spheres and one 4 mm glass sphere. Samples were kept on ice throughout. The samples were subject to bead beating using a FastPrep-24 bead beater (MP Biomedical) at a speed of (6.0 m/s) for 30s with a period of 5 min on ice between each beating. To determine optimal number of bead beating steps this process was repeated up to 6 times. The bead beating step was repeated a further 2 times for optimal recovery of intracellular protein. Samples were subject to centrifugation at 20,000 x g for 30 min at 4 °C and the supernatant from this step was filtered through a 100K Amicon Ultra centrifugal filter tubes (Millipore, Darmstadt, Germany) according to the manufacturer's instructions. For the extracellular only samples, this centrifugation step was conducted immediately instead of the bead beating step. Supernatant after filtration was transferred to new sterile tubes and taken as the crude protein extract. NaN_3 was added aseptically to each sample to a final concentration of 0.05% w/v neat samples,10-fold and 100-fold dilution were used to estimate protein concentration using the bicinchononic acid assay (BCA) method according to the manufacturer's instructions (PIERCE, Rockford, IL, USA) and samples were normalised to 1 mg/mL protein using the appropriate buffer as a diluent to conduct subsequent protease activity estimates. Azo-casein assay were performed as described in General Methods (Chapter 2.0 section 2.9.1). Samples were kept in the freezer at -20 °C. The protein concentration measurements and protease activity estimates were performed on the same samples after 24 hours, 1 week, 1 month, 3 months, 6 months and 1 year.

6.4.5 Design, conduct and analysis of the effects of storage of whole faecal samples at -20 and -80 °C:

To assess whether long term storage of faecal samples provides impacts on the reproducibility on measurements of protein concentration and protease activity, faecal samples from 3 healthy volunteers were collected. Each sample was thoroughly mixed in a sterile environment and 13 lots of 1 g specimens were separated into sterile containers. 1 sample from each individual was processed immediately, and 6 samples from each individual were stored at -20 ˚C and the remaining 6 were stored at -80 ˚C. The fresh sample was subject to crude protein extraction, protein measurements and protease assay as previously described. The frozen samples were analysed in exactly the same way after 6 storage time points; 24 hours, 1 week, 1 month, 3 months, 6 months and 1 year.

6.4.6 Statistical Analysis

Differences between the buffers and differences between the storage of faecal samples at sub-zero temperatures in reproducibility of mean protein concentration was determined by correlation analysis using the spearman method and the pairwise comparison using the students t-test using a holm adjustment for familywise error after the samples had been demonstrated to follow a normal distribution. The same statistical analysis was implemented to compare the mean protease activity over time.

To determine optimal bead beating steps mean enzyme activities were compared with pairwise comparisons using the Wilcoxon signed rank test with a Holm adjustment method. Similarly to determine the effect of NaN_3 mean enzyme activity was compared using the Wilcoxon signed rank test. All statistical analysis was conducted in R software.

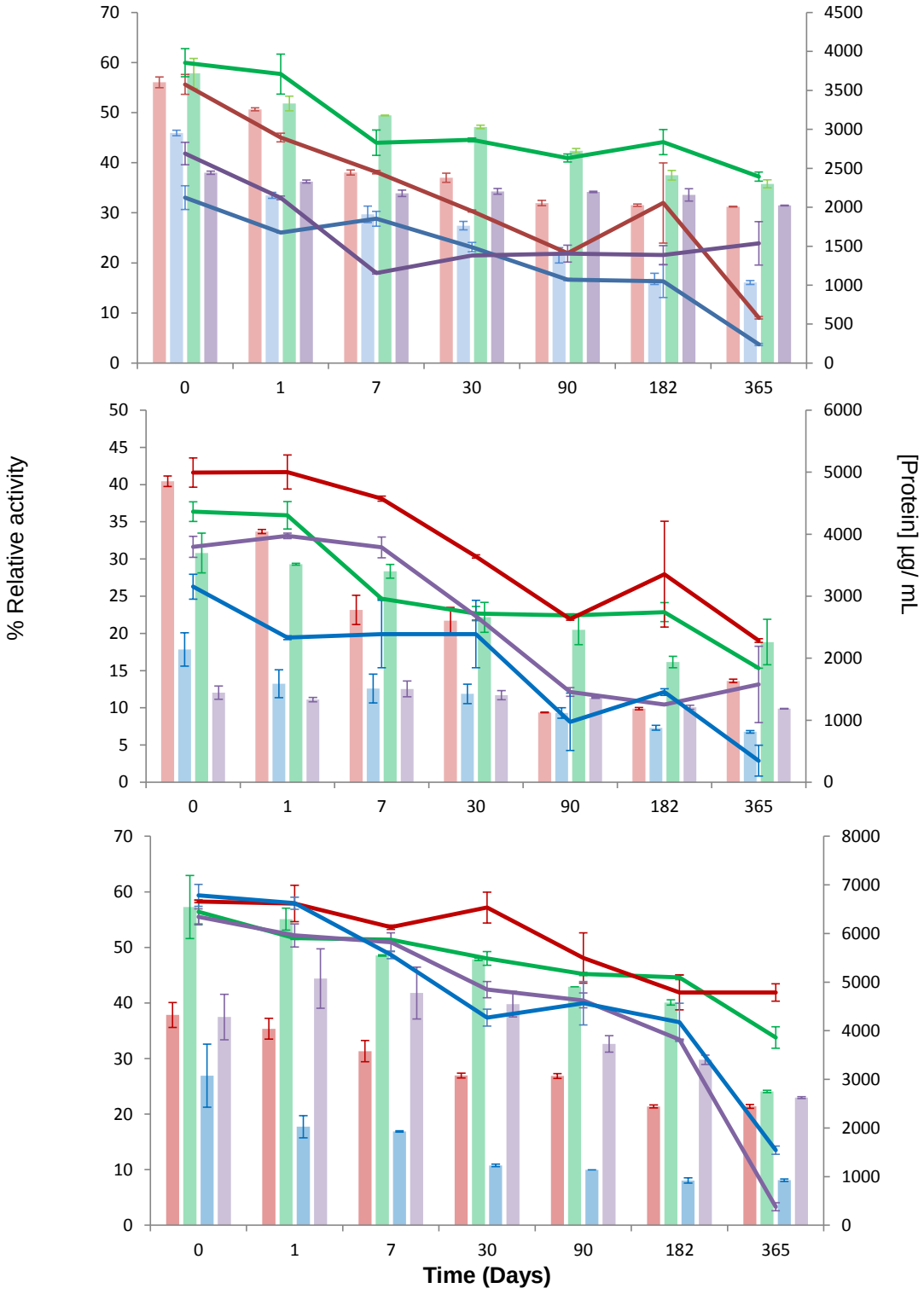
6.5 REFERENCES

1. Gill, S.R., Pop, M., DeBoy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., Gordon, J.I., Relman, D.A., Fraser-Liggett, C.M., and Nelson, K.E. (2006). Metagenomic Analysis of the Human Distal Gut Microbiome. Science *312*, 1355-1359.
2. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59-65.
3. Gloux, K., Berteau, O., El Oumami, H., Beguet, F., Leclerc, M., and Doré, J. (2011). A metagenomic beta-glucuronidase uncovers a core adaptive function of the human intestinal microbiome. Proceedings of the Nationall Academy of Science U S A *108 Suppl 1*, 4539-4546.
4. Jones, B.V., Begley, M.i., Hill, C., Gahan, C.G.M., and Marchesi, J.R. (2008). Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. Proceedings of the National Academy of Sciences *105*, 13580-13585.
5. Flores, R., Shi, J., Gail, M.H., Ravel, J., and Goedert, J.J. (2012). Assessment of the human faecal microbiota: I. Measurement and reproducibility of selected enzymatic activities. European Journal of Clinical Investigation. 2012 Aug;42(8):848-854. doi: 10.1111/j.1365-2362.2012.02660.x. Epub 2012 Mar 13.
6. Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S.R., Nelson, K.E., and Relman, D.A. (2005). Diversity of the human intestinal microbial flora. Science. 2005 Jun 10;308(5728):1635-8. Epub 2005 Apr 14.
7. Peterson, D.A., Frank, D.N., Pace, N.R., and Gordon, J.I. (2008). Metagenomic Approaches for Defining the Pathogenesis of Inflammatory Bowel Diseases. Cell Host & Microbe *3*, 417-427.
8. Ekkers, D.M., Cretoiu, M.S., Kielak, A.M., and Elsas, J.D. (2012). The great screen anomaly--a new frontier in product discovery through functional metagenomics. Applied Microbiology and Biotechnology *93*, 1005-1020.
9. Cowan, D., Meyer, Q., Stafford, W., Muyanga, S., Cameron, R., and Wittwer, P. (2005). Metagenomic gene discovery: past, present and future. Trends in Biotechnology *23*, 321-329.
10. Bertrand, H.l.n., Poly, F., Van, V.T., Lombard, N., Nalin, R., Vogel, T.M., and Simonet, P. (2005). High molecular weight DNA recovery from soils prerequisite for biotechnological metagenomic library construction. Journal of Microbiological Methods *62*, 1-11.
11. Salonen, A., Nikkilä, J., Jalanka-Tuovinen, J., Immonen, O., Rajilić-Stojanović, M., Kekkonen, R.A., Palva, A., and de Vos, W.M. (2010). Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: Effective recovery of bacterial and archaeal DNA using mechanical cell lysis. Journal of Microbiological Methods *81*, 127-134.
12. Purohit, M.K., and Singh, S.P. (2009). Assessment of various methods for extraction of metagenomic DNA from saline habitats of coastal Gujarat (India) to explore molecular diversity. Letters in Applied Microbiology *49*, 338-344.
13. Yuan, S., Cohen, D.B., Ravel, J., Abdo, Z., and Forney, L.J. (2012). Evaluation of methods for the extraction and purification of DNA from the human microbiome. PLoS One *7*.

14. McOrist, A.L., Jackson, M., and Bird, A.R. (2002). A comparison of five methods for extraction of bacterial DNA from human faecal samples. Journal of Microbiological Methods *50*, 131-139.
15. Salonen, A., Nikkila, J., Jalanka-Tuovinen, J., Immonen, O., Rajilic-Stojanovic, M., Kekkonen, R.A., Palva, A., and de Vos, W.M.(2010). Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. Journal of Microbiological Methods *81*, 127-134.
16. Lauber, C.L., Zhou, N., Gordon, J.I., Knight, R., and Fierer, N. (2010). Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. FEMS Microbiology Letters *307*, 80-86.
17. Carroll, I.M., Ringel-Kulka, T., Siddle, J.P., Klaenhammer, T.R., and Ringel, Y. (2012). Characterization of the fecal microbiota using high-throughput sequencing reveals a stable microbial community during storage. PLoS One *7*, e46953.
18. Macfarlane, G.T., Cummings, J.H., and Allison, C. (1986). Protein Degradation by Human Intestinal Bacteria. J Gen Microbiol *132*, 1647-1656.
19. Lacerda, C.M.R., and Reardon, K.F. (2009). Environmental proteomics: applications of proteome profiling in environmental microbiology and biotechnology. Briefings in Functional Genomics & Proteomics *8*, 75-87.
20. Kolmeder, C.A., de Been, M., Nikkilä, J., Ritamo, I., Mättö, J., Valmu, L., Salojärvi, J., Palva, A., Salonen, A., and de Vos, W.M. (2012). Comparative Metaproteomics and Diversity Analysis of Human Intestinal Microbiota Testifies for Its Temporal Stability and Expression of Core Functions. PLoS ONE *7*, e29913.
21. Wilmes, P., and Bond, P.L. (2006). Metaproteomics: studying functional gene expression in microbial ecosystems. Trends in Microbiology *14*, 92-97.
22. Costa, F., Mumolo, M.G., Ceccarelli, L., Bellini, M., Romano, M.R., Sterpi, C., Ricchiuti, A., Marchi, S., and Bottai, M. (2005). Calprotectin is a stronger predictive marker of relapse in ulcerative colitis than in Crohn's disease. Gut *54*, 364-368.
23. Chassaing, B., Srinivasan, G., Delgado, M.A., Young, A.N., Gewirtz, A.T., and Vijay-Kumar, M. (2012). Fecal lipocalin 2, a sensitive and broadly dynamic non-invasive biomarker for intestinal inflammation. PLoS One *7*, e44328.
24. Möstl, E., and Palme, R. (2002). Hormones as indicators of stress. Domestic Animal Endocrinology *23*, 67-74.
25. Freeman, H.J. (1986). Effects of Differing Purified Cellulose, Pectin, and Hemicellulose Fiber Diets on Fecal Enzymes in 1,2-Dimethylhydrazine-induced Rat Colon Carcinogenesis. Cancer Research *46*, 5529-5532.
26. Gibson, S.A., McFarlan, C., Hay, S., and MacFarlane, G.T. (1989). Significance of microflora in proteolysis in the colon. Appl Environ Microbiol *55*, 679-683.
27. Macfarlane, G.T., and Macfarlane, S. (2012). Bacteria, colonic fermentation, and gastrointestinal health. Journal of AOAC International *95*, 50-60.
28. Steck, N., Mueller, K., Schemann, M., and Haller, D. (2013). Bacterial proteases in IBD and IBS. Gut.
29. Hughes, R., Magee, E.A., and Bingham, S. (2000). Protein degradation in the large intestine: relevance to colorectal cancer. Current Issues in Intestigative Microbiology *1*, 51-58.
30. Flintoft, L. (2003). A bad influence. Nature Reviews Cancer *3*, 890-890.

31. Ulger Toprak, N., Yagci, A., Gulluoglu, B.M., Akin, M.L., Demirkalem, P., Celenk, T., and Soyletir, G. (2006). A possible role of Bacteroides fragilis enterotoxin in the aetiology of colorectal cancer. Clinical Microbiology and Infection *12*, 782-786.
32. Shen, H.-B., and Chou, K.-C. (2009). Identification of proteases and their types. Analytical Biochemistry *385*, 153-160.
33. Rawlings, N.D., Barrett, A.J., and Bateman, A. (2014). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Research *40*, D343-D350.
34. Steck, N., Hoffmann, M., Sava, I.G., Kim, S.C., Hahne, H., Tonkonogy, S.L., Mair, K., Krueger, D., Pruteanu, M., Shanahan, F., et al. (2011). *Enterococcus faecalis* metalloprotease compromises epithelial barrier and contributes to intestinal inflammation. Gastroenterology *141*, 959-971.
35. Wu, S., Rhee, K.-J., Zhang, M., Franco, A., and Sears, C.L. (2007). *Bacteroides fragilis* toxin stimulates intestinal epithelial cell shedding and gamma-secretase-dependent E-cadherin cleavage. Journal of Cell Science *120*, 1944-1952.
36. Wu, S., Lim, K.-C., Huang, J., Saidi, R.F., and Sears, C.L. (1998). *Bacteroides fragilis* enterotoxin cleaves the zonula adherens protein, E-cadherin. Proceedings of the National Academy of Sciences *95*, 14979-14984.
37. Lozupone, C.A., Stombaugh, J.I., Gordon, J.I., Jansson, J.K., and Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. Nature *489*, 220-230.
38. Flores, R., Shi, J., Gail, M.H., Gajer, P., Ravel, J., and Goedert, J.J. (2012). Association of Fecal Microbial Diversity and Taxonomy with Selected Enzymatic Functions. PLoS ONE *7*, e39745.
39. Smith, P.K., Krohn, R.I., Hermanson, G.T., Mallia, A.K., Gartner, F.H., Provenzano, M.D., Fujimoto, E.K., Goeke, N.M., Olson, B.J., and Klenk, D.C. (1985). Measurement of protein using bicinchoninic acid. Analytical Biochemistry *150*, 76-85.
40. Privalov, P,L(1990). Cold Denaturation of Protein. Critical Reviews in Biochemistry and Molecular Biology *25*, 281-306.
41. Khan, M.Z., Altmann, J., Isani, S.S., and Yu, J. (2002). A matter of time: evaluating the storage of fecal samples for steroid analysis. General and Comparative Endocrinology *128*, 57-64.

SUPPLEMENTARY MATERIALS



←*from previous page*

Supplementary materials figure 6 3. Protein concentration estimates and protease activity assay results over time for 3 individuals.

The line graph shows the relative protease activity over time (primary y axis) and the bar graph represents protein concentration (secondary y axis) over time for each individual. Red data indicate total protein extraction stored in buffer 1, blue data indicate extracellular protein stored in buffer 1, green data indicate total protein extracts stored in buffer 2 and purple data indicate extracellular protein stored in buffer 2. Samples are plotted as the average of triplicate measurements for both protein concentration estimates and protease activity assay and the error bars represent the standard deviation.

7.0 A COMPREHENSIVE COMPARISON OF THE HEALTHY AND IBD FEACAL DEGRADOME

7.1 INTRODUCTION

Inflammatory bowel disorders (IBD) encompass a spectrum of diseases of which the two main pathologies are; Ulcerative colitis (UC) and Crohn's disease (CD). They are debilitating and chronically recurring disorders of the human gastrointestinal tract characterised by chronic inflammation of the intestinal mucosa [1-3] and an inability to control an inflammatory response to an unknown stimulus [3]. The onset of the disease occurs predominantly in individuals between the ages of 15-30 years [1] mainly in inhabitants of the developed world.

Symptoms of IBD include severe abdominal pain and discomfort, bloating, periods of constipation, periods of diarrhoea, a highly disturbed bowel pattern, gastrointestinal bleeding and weight loss [3, 4]. Research has also suggested a correlation between IBD and psychiatric disorders such as anxiety and depression.[5] Evidently, the disease has a profound impact on people's lives especially their quality of life, its occurrence is also increasing and as a result, further consideration must be given to the increasing burden on the healthcare system [6] and other socioeconomic indices.

Symptoms of CD and UC differ in that UC is characterised pathologically by ulceration of the mucosa and is limited to the rectum and the colon. Patients suffering often have mucus-like and bloody diarrhoea while CD can affect any region of the gastrointestinal tract and is characterised by diarrhoea, anorexia, weight loss and involves transmural rather than mucosal inflammation. Therapeutics are administered to ameliorate the symptoms of the disease and treatments often include the prescription of immunosuppressive drugs such as Azathiaprine, steroids; commonly prednisolone and anti-inflammatory drugs such as Pentasa and Asacol, and in more severe cases, surgical removal of affected regions of the gut including abscesses and fistulas may be required to alleviate pain. However, IBD remains irremediable and is chronically relapsing. Removal of fistulas and abscesses by surgical intervention do not cure the disease, they are a means of alleviating pain in an attempt to improve the patients quality of life [7].

The cause of IBD is complex and remains idiopathic as it has been shown that there are a number of factors, interacting with one another that lead to the onset of the disease. The disease can be widely attributed to; genetics factors, mucosal immunity and interactions of the gut microbiota, and is often attributed to the initiation of an inflammatory response by the host's impaired immune system to commensal gut bacteria. Figure 7.1 highlights the main factors contributing to IBD for which the factors that have received the most established and more convincing evidence for involvement in IBD will be discussed.

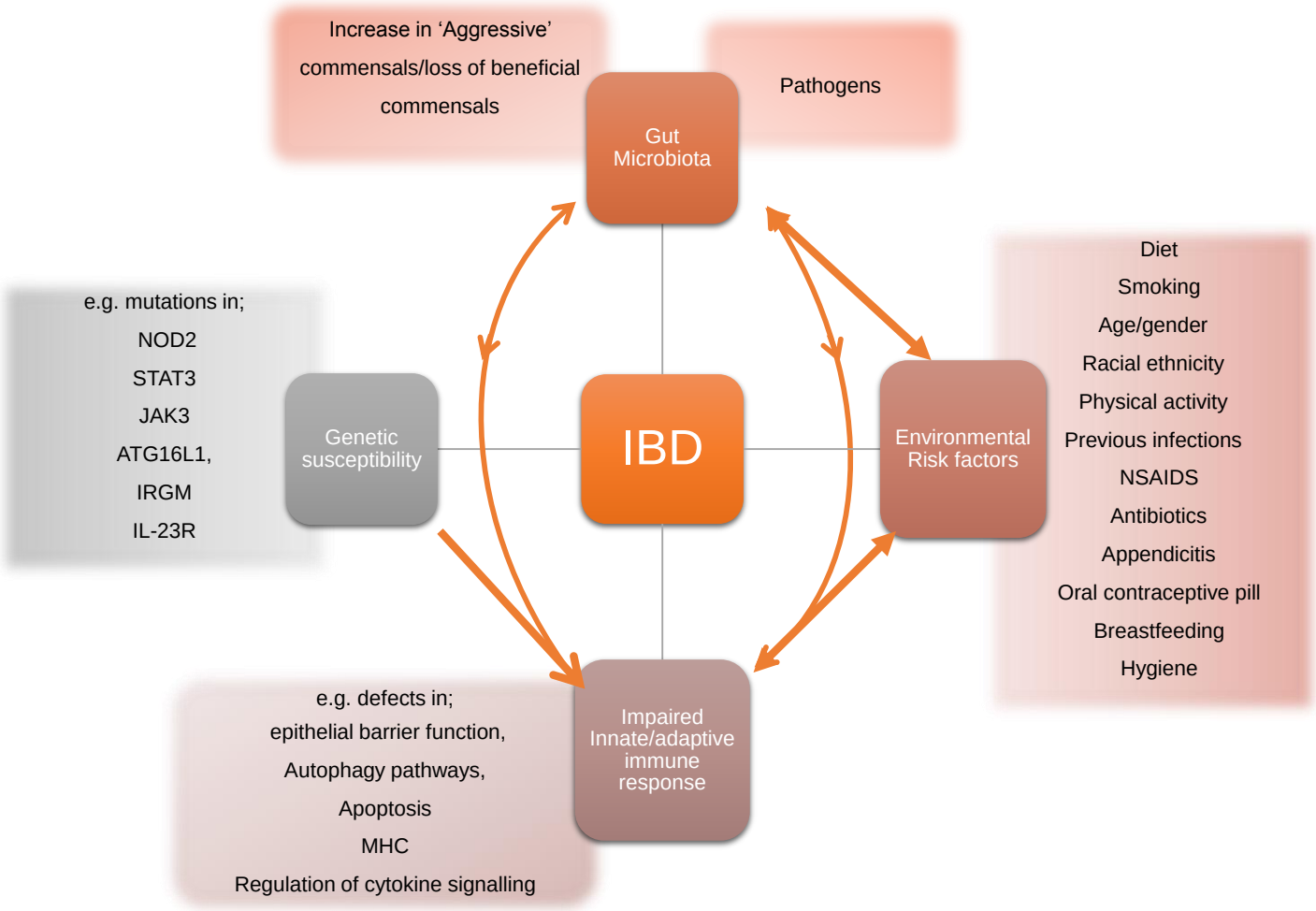


Figure 7.11. Contributory factors to IBD and their interactions. The pathogenesis of IBD cannot be attributed to a solitary factor; in fact it is currently thought that the disease is a consequence of multiple factors including interplay between an impaired immune system, the gut microbiota and environmental factors with an underlying genetic predisposition for disease development.

IBD is generally considered a disease of the developed world and consequently, environmental stimuli have also been implicated with IBD, namely smoking, diet, stress, and the use of certain drugs [8].

The relationship between IBD and smoking is complex and studies have shown a contrasting correlation between the two main disease pathologies. Smoking has been found to be a significant risk factor in CD but UC is more associated with ex-smokers or non-smokers and has even been considered to be a beneficial environmental factor in preventing disease progression [9]. The mechanisms by which smoking is hypothesised to influence IBD include; modulating cellular immunity[10], altering gut permeability and motility [11], varying mucosal blood flow, altering mucus production and stimulating the release of glucocorticoids [12, 13]. It is conjectured that nicotine is the metabolite responsible for the observed effects of smoking in IBD as nicotinic acetylcholine receptors (nAChRs) are present in cells of the colonic epithelium [14] and T-cells have been found to express nAChRs suggesting that nicotine may be able to modulate T-cell activity [15]. However there remains no clear evidence for this to date and the results of clinical trials remain modest [12].

The relationship between diet and IBD is similarly complex and has been the subject of much research given that IBD is most prevalent in the western world. Studying the effect of diet on disease progression presents a challenge as in human subject studies it difficult to determine the component of each diet that may be influencing the gut mucosa and the compliance of an individual to accurately describe their diets is another limiting factor that must be considered. In a Japanese cohort study, it was found that the consumption of sugars, sweeteners, fats and oils were positively associated with IBD [16] similar findings regarding sugar intake have been observed in other population based studies in the USA, Canada [17] and Europe [18-20]. Conversely, a diet high in fibre, fruit and vegetables has been shown to be negatively correlated with IBD [20, 21]. However, many studies on the dietary impact on IBD have reported inconclusive findings and there are significant challenges to effectively study the role of diet [8].

Stress has also been implicated in the pathogenesis of IBD largely due to the effect it can have on the immune system function [22]. However the specific pathways by which psychological stress may be involved in IBD are again, unknown. Animal models have

provided some insight into how chronic stress exacerbates a colitis phenotype by impairing the immune system, increase intestinal permeability, impairing mucosal epithelia barrier integrity and defence against bacteria [23, 24].

There are two main drug groups that have received attention for their association with IBD and they are the oral contraceptive pills (OCP) and non-steroidal anti-inflammatory drugs (NSAIDs). A meta-analysis of a putative role of the OCP in IBD deduced a modest positive relationship in women who took the oral contraceptive and development of IBD (both UC and CD) compared to controls, however the associations found were weak and limited to small data sets [25]. The basis for OCP contribution to IBD aetiology stems from the effects of oestrogen [26] and its association with multifocal, gastrointestinal infarcation and putative thrombogenic effects [27].

NSAIDs are the other leading class of drugs that have received significant attention for implication in IBD and there is far less ambiguity surrounding the involvement of these drugs. Studies have shown that usage of NSAIDs is positively correlated with the development or exacerbation of IBD [28] The mechanism of action proposed for how they are implicated in IBD differs to that of OCP. NSAIDs have been shown to cause significant damage to the upper gastrointestinal tract, the small intestine and the colon [29] [28] by enhancing epithelial permeability, inhibiting cyclooxygenase 2 (COX2) and enterohepatic recirculation [29]. More specifically to IBD are NSAIDs role in inhibiting COX2, a regulatory enzyme in the biosynthetic pathway of prostaglandins (PG) [30] which, as well as having many other critical cellular functions, also function in mediating intestinal epithelial inflammation [31] particularly by initiating anti-inflammatory cytokines namely interleukin (Il)-10. Studies on Il-10 deficient mouse models have shown that colitis develops faster, and with increased severity and consequently it has been found that functioning PG are integral in the inhibition of chronic inflammation [8, 30]. Coupling this information with other studies showing that NSAIDs are correlated with the exacerbation or remission of IBD [28] it has become widely accepted that usage of NSAIDs in IBD patients should be avoided if possible.

There is mounting evidence to suggest a role for certain environmental stimuli as risk factors in the development or exacerbation of IBD, and in some cases, in the protection against the disease. However, current data is still very limited and there is a level of

ambiguity surrounding some of the findings. In order to determine which environmental factors are integral to the onset of the disease more studies are required to deduce the molecular basis of their interaction with the gut and how they might be affecting mucosal immunity and gut function and consequently causing inflammation.

Population studies investigating the occurrence of IBD amongst relatives have suggested that genetically determined factors may be contributing to the disease. It has been found that there is up to a 10-fold greater risk of individuals within families of developing IBD [32]. Further evidence for the heritability of the disease has been generated by the studies of monozygotic twins compared to dizygotic twins whereby the rates of monozygotic twins developing the disease are much higher than in dizygotic twins(10-15%) [33]. It is now widely accepted that genetics factors are associated with IBD. The implementation of genome wide association studies (GWAS) and deep-sequencing analyses have enabled the identification of a number of susceptibility loci for IBD. It has become clear that IBD cannot be attributed to one gene alteration but is rather a heterogeneous polygenic disease [1] and variations at a number of different loci render susceptibility to IBD rather than definite onset of the disease as a result. Over 100 susceptibility loci have been identified in large cohorts of patients with IBD [34-50].

Gene identification and their encoded proteins enable the analysis of the molecular mechanism and function of these genes and how gene variants might be a contributory factor in IBD. The majority of the well-characterised loci have been found to be genes that encode proteins involved in both the innate and adaptive immune system, the autophagy pathway and in maintaining epithelial barrier integrity and function. Despite CD and UC conferring different clinical features, GWAS studies have shown concordance in the susceptibility loci between the two suggesting that each disease phenotype shares common pathways [51].
One of the first well characterised loci to be discovered and implicated in IBD is that of the pericentometric region of chromosome 16 called *IBD-1*.Within this region is a gene called *NOD2* (nucleotide-binding oligomerization domain containing 2 or sometimes known as *CARD15*) of which a cytosine insertion mutation in this gene is associated with Crohn's disease [38]. *NOD2* is a pattern recognition receptor protein and the wild type protein activates nuclear factor kappa B (NF-κB) which regulate the expression of a number of genes involved in both innate and adaptive immunity in response to bacterial

lipopolysaccharides. NF-kβ regulates the expression of a number of genes in both innate and adaptive immunity including interleukins (IL) 1,2,6,8,9,10,11,12,13,15, tumour necrosis factor (TNF) α and β, interferons (IFN) β and y [52]. Mutations in *NOD2* therefore render the host less able to respond to a bacterial stimulus.

Other identified risk loci that form part of the adaptive immune system and their regulation of interleukin receptor pathways include regions containing the genes; *STAT3, JAK3*, and *TYK2*, *IL-23R* and *IL12B*. Signal transducer and activator of transcription 3 (*STAT3*), janus kinase 2(*JAK2*), a tyrosine kinase and tyrosine kinase 2 (*TYK2*) are involved in a fundamental cellular signalling process and coordinate the transcription of immunregulatory components. Mutations in this pathway again mean the host is deficient in forming an appropriate response to a particular stimulus. Genetic variants of the genes encoding interluekins 12, 17 and 23 have also been identified as risk factors for CD and UC and their expression has been found to be induced in active forms of IBD [53] suggesting their role as proinflammatory mediators if dysregulated. Some alternative variants of IL-23 receptor genes have also been found to confer protection against IBD [54].

Another common pathway associated with genetic susceptibility in IBD is related to that of autophagy. Some of the genes identified via GWAS studies involved in autophagy include *ATG16L1*, *IRGM* and *LRRK2*. Autophagy is an integral cell homeostasis mechanism involved in the degradation and recycling of intracellular components and organelles and the targeted degradation of intracellular pathogens. *ATG16L1* (Autophagy related protein 16-1) as its name suggests, is an essential protein for the autophagy process. Variants of this gene that lead to loss in function e.g. the single nucleotide polymorphisms (SNPs) rs2241880 and polymorphism Thr300Ala have been strongly associated with increased risk for CD [55]. SNPs in the immunity-related GTPase family M gene (*IRGM*) leading to loss of its wildtype function of the interferon induced autophagy pathways targeting intracellular pathogens and has been strongly related to both CD and UC [56]. Similarly, the Leucine-rich repeat kinase 2 (*LRRK2*) enzyme is also involved in the intracellular degradation of pathogens and mutations lead to a deficiency in this process and an exacerbated colitis has been observed in *LRRKS2* deficient mouse models [57].

The final pathway discussed here that's heavily implicated in both forms of IBD involve genes involved the structure and function of the gut epithelia. Aside from its fundamental role in nutrient absorption, the gut epithelia acts as both a physical and biochemical barrier protecting the underlying tissue from not only pathogenic bacteria but also preventing the helpful commensal bacteria from infiltrating the underlying tissue. It plays an active role in sensing and responding to microbial stimuli to illicit an appropriate immune response to maintain gut homeostasis. Dysregualtion of epithelial barrier integrity can have severe consequences for host health due to an enhanced probability of bacteria and bacterial molecules translocating and is therefore implicated with the development of IBD. Some of the better characterised genes involved in the causation of a dysregulated epithelia barrier include *CDH1*, *GNA12*, *PTPN2*, *HNF4A*, *LAMB1*, *OCTN2* and *ECM1*. For example, *CDH1* encodes epithelial E-cadherin, a transmembreane glycoprotein involved in epithelial cellular adhesion. Truncated forms of this protein and a consequent increase in permeability have been associated with CD [58] as have a number of the other genes mentioned that encode proteins that are involved in cellular junction integrity and protection against epithelial permeability [59, 60].

Unsurprisingly given the vast number of antigenic stimuli in the gut from bacteria, food and drink etc. and because of their intrinsic involvement in cytokine secretion and regulation of the inflammatory process, dysregulation of various cells of the gut-associated lymphoid tissue (GALT) such as T and B cells have also been implicated in the pathogenesis of IBD. For example T-cells, which secrete cytokines as part of the inflammatory process in response to antigens processed by antigen presenting cells (APCs) have been demonstrated to have an abnormal response to 'normal' antigens i.e. resident commensal bacteria [61]. Abnormal levels of circulating CD4(+) (cytokine secreting) and CD8(+) T cells (cytotoxic) have also been found in IBD patients when compared with healthy volunteers suggesting a role for over-activation of T-cells in the progression of IBD [62].

Studies of the pathogenesis of IBD have been hugely advanced with the use of experimental animal models, namely murine models with altered epithelial barrier function and innate or adaptive immune responses [63]. The selective disruption of genes such as those mentioned above are commonly implemented to study IBD for

example disruption of STAT3 in macrophages results in Il-10 deficient mice and has been used to demonstrate Il-10 deficient mice significantly increase inflammatory cytokine production including TNFα, Il-1, IFNy and IL-6 [64]. Other models directly target the epithelial barrier e.g. Toll-like receptor 4 (TLR4) and myeloid differentiation primary response gene (MyD88) deficient mice subsequently treated with Dextran sulphate sodium (DSS), which cause acute colitis by disrupting the mucosal barrier have shown TLR4 to be involved in prevention of bacterial translocation following epithelial injury [65]. Increasingly sophisticated animal models are now being utilised to study specific questions regarding the aetiology of IBD and have highlighted that functional defects in the epithelial barrier, roles of numerous cytokines, T-cells and many other cells and proteins of both the innate and adaptive immune response are contributing to IBD however, perhaps one of the most significant findings has been that despite a clear, yet complex interplay between numerous genetic loci, the actual development of colitis is dependent upon the presence of the gut microbiota [63].

The fundamental question of which bacteria are involved and more specifically what molecular mechanisms of these bacteria are contributing to the progression of colitis are yet to be solved and remain a significant factor in determining the pathophysiology of this complex disease.

Recent deep-sequencing approaches and the implementation of powerful bioinformatics tools [66, 67] have enabled us to recognise the genetic diversity and potential of the gut microbiota and have allowed us to elucidate that dysbiosis of the gut microbiota is associated with IBD [68]. One of the most frequently observed general dysbiosis is the marked reduction in the *Firmictues* phylum and an increase in members of the *Actinobacteria*. It has been considered that the changes in composition and diversity of the gut microbiota can lead to enrichment of pathogenic bacteria which are contributing to initiation of the disease and parallel to this, a decrease in ‘protective’ commensals and their products may also be a contributory factor. Other evidence for a role of bacteria in IBD include; the loss of epithelial barrier integrity and increased bacterial translocation, an insufficient or overactive response to bacterial stimuli, dysbiosis leads to commensal with altered metabolic activities which may be harmful to the host or lastly, the presence of a specific pathogen.

Research on the community of gut microbiota in IBD are generally commensurate in that the disease can be characterised by a dysbiosis of the gut microbiota, however the question remains whether this dysbiosis is a cause or an effect of the disease.
Patients with IBD often have extraintestinal complications, increased risk of developing CRC. However, regardless of whether dysbiosis is cause or effect, the fact remains that the microbiota is altered. It is vital that the key players and the differences in molecular, metabolic and enzymatic activities as a result of dysbiosis are determined for a number of reasons; a) this may be a putative cause of the disease, b) to deduce the potential long term effects of dysbiosis and c) if dysbiosis is a consequence, understanding the dominant organisms and their metabolic capacity may help us determine the driving force behind these alterations.
Substantial research has been conducted into determining the presence of a specific pathogen as a solitary causative agent of IBD. *Mycobacterium avium* subspecies *paratuberculosis* (MAP), adherent-invasive *E. coli* (AIEC) *Helicobacter* spp. and *Campylobacter* spp. have received the most attention for their putative role in the disease. MAP first received attention for implications with IBD due to its similarity with the bovine condition Johne's disease in which MAP is also the causative agent. It is also commonly isolated from meat, dairy and sometimes water so there are numerous sources of infection [69]. MAP has been shown to induce intestinal inflammation and severe mucosal damage via goblet cell invasion [69]. However while in some cases MAP has been isolated from a significant majority of IBD patients [70] other studies have failed to isolate the bacteria in cohorts of IBD patients [71] and so the involvement of MAP in IBD remains controversial.

AIEC has been frequently isolated from CD patients and has been shown to adhere and invade intestinal epithelial cells by binding to adhesion receptor carcinoembryonic antigen-related cell adhesion molecule 6 (CEACAM6) with type one pili and flagella [72] and recruiting actin microfilaments and microtubules. They can pass the mucosal barrier by way of a long polar fimbriae interacting with the Peyer's Patches of the epithelial lining [73] and from here, can survive and replicate in macrophages without inducing cell death. Numerous studies have also shown that AIEC can induce colitis in genetically susceptible animal models [74]. AIEC has been significantly associated with ileal crohns lesion tissue and there is increasingly strong evidence for its involvement in

the disease (36.4% compared to 6.2% in controls) however, it is rarely isolated from UC patients [74].

Helicobacter and *Campylobacter* sp. were originally suggested to be implicated with IBD due to their previously concluded causation of gastric ulcers and gastroenteritis respectively. Interestingly, *Helicobacter pylori* prevalence has commonly been found to be significantly lower in IBD patients than in healthy [75, 76]. Other *Helicobacter* species including *H. hepaticus, H. bilis, H. trogontum, H. rodentium* and *H. typhlonius* however, have shown the ability to induce colitis in experimental animal models [77, 78] and have been positively associated with UC [79]. The case is similar for non-*jejuni Camplyobacter* species too. Species implicated in IBD, namely CD include *C. concisus, C. hominis, C. ureolyticus* and *C. gracilis* [80]. *C. concisus* has perhaps received the most attention and has been found to be significantly more prevalent in children with CD compared to healthy controls [81]. *C concisus* has been shown to attach to epithelial cells and disrupt epithelial junction proteins causing a loss in barrier integrity and a consequent increase in epithelial permeability [82].
Clearly there is much supporting evidence for the theory of a specific pathogen, however rarely are they found in all sufferers of CD or UC and even more rarely do they appear in both forms of the disease nor do any of the mentioned pathogens meet all four criteria of Koch's postulates. One feature that becomes apparent when studying pathogens and indeed commensals is that many of them have the ability to become invasive and have stimuli that induce a pro-inflammatory response. The primary molecules responsible for this are virulence factors [83].

Efforts are continuously being made to identify the molecular structures and mechanisms of pathogenic or commensal bacteria that may be acting as virulence factors and are responsible for their contribution to the pathogenesis of IBD. Virulence factors that have received particular attention with regards to their role in IBD include adhesins, bioilm formation, [84] motility, type-specific secretion systems and exotoxins [85].
Aside from these factors, many of the molecular features of IBD such as alteration in innate immune defence, chronic inflammation, impaired intestinal epithelial barrier function can be associated with a well-established virulence factor of microorganisms and that is; protease activity and the term 'degradome' is used to encompass the entire

proteolytic repertoire of single microorganisms or the metagenome from a particular environment (which the case of this research will be the human gut).

Microbial proteases have been proven to be responsible in part or even entirely for the virulence of certain microorganisms and the onset of many diseases and there is increasing evidence implicating them in the pathology of diseases of the human gut (see Table 7.1). Host proteases have also been implicated in IBD; a large scale study incorporating a number of European studies analysing the ancestry of IBD were used by the IBDase consortium (an FP7 project now completed) to screen for protease and protease inhibitor genes in critical regions inferring susceptibility to IBD. This study identified 75 protease genes and 7 protease inhibitor genes in CD and 14 protease and 4 protease inhibitor genes in UC [86] as well as matrix metalloproteases (MMP) that had previously been implicated in IBD.

Table 7.1. A Description of proteases produced by pathogens and opportunistic pathogens of the human gastrointestinal tract (inclusive of the oral microbiota) and their host target.

MMPs are zinc-dependent endopeptidases the activity of which has been associated with IBD via the uncontrolled degradation of extracellular matrix (ECM), a major feature of IBD, and subsequent tissue damage. MMP activity has been shown to be significantly

Species	Classification	Protease	Host target	Ref
Bacteroides fragilis	Opportunistic pathogen	Metalloprotease; fragylisin	Degradation of E-cadherin via induction of γ-secretase dependent shedding and subsequent loss of epithelial barreire function	[87]
Enterohaemorrhagic Escherichia coli	Pathogen	Metalloproteinase *StcE*	Cleavage of mucin 7 and glycoprotein 340, facilitation of adherence to epithelial cells	[88]
Enterococcus faecalis	Opportunistic pathogen	Metalloprotease; Gelatinase E (*GelE*)	Degradation of extracellular domain of E-cadherin and subsequent loss of epithelial barrier function	[89]
Helicobacter pylori	Pathogen	Serine protease high temperature requirement A (HpHtrA)	Facilitates the degradation of E-cadherin	[90]
Campylobacter jejuni	Pathogen	*HtrA* homologue	Aids with adherence and invasion in a INT407 cell line	[91]
Shigella flexneri	Pathogen	Pic serine protease	Facilitates bacterial adherence and invasion by degradation of mucin. Also has haemagglutinin activity	[92]
Clostridium difficile	Pathogen	Cwp84 cysteine protease	Degradation of ECM	[93]
Porphyromonas gingivalis	Pathogen	Cysteine protease Gingipain	Degradation of IL-1β, Il-6 and TNFα and can truncate IL-8 resulting in enhanced neutrophil recruitment	[94]
Listeria monocytogenes	Pathogen	ClpC Serine protease	Promotes escape from the phagosmome of macrophages and aids with adherence and invasion	[95]
Vibrio cholerae	Pathogen	Metalloprotease; vibriolysin/HA protease	Morphological changes in filamentous actin and Zonula occludens-associated protein (ZO-1); alters epithelial barrier function.	[96]

elevated in the inflamed mucosa compared to non-inflamed with a particular increase in MMP-9 [97] . Other MMPs such as MMP 1 (a collagenase) and MMP-3 (stromelysin-1) have been found to be unregulated in patients with IBD, while their endogenous inhibitors; Tissue inhibitor of metalloproteases (TIMPs) have been found to be in insufficient quantities to control the level of MMP in inflamed areas of the gut [98]. As a result of such findings, studies have focused on the effects of metalloprotease inhibitors on colitis. Through the use of animal models, many studies have concluded that administration of MMP inhibitors such as Marismastat [99], Batimastat[100] and Phenantroline [101] in animal models of colitis have significantly reduced the effects of inflammation.

Elafin, an elastase specific natural serine protease inhibitor has also been found to be reduced in IBD patients [102] and is associated with an increase in elastolytic activity. It has also been demonstrated that the delivery of elafin to the colon, via elafin producing strains of lactic acid bacteria, confers protection against colitis, restoration of epithelial homeostasis as well as a decrease in elastolytic activity in the colon [102].

Evidently, protease activity represents a good therapeutic target for IBD, and this clearly demonstrates that they are involved in disease aetiology. Though since MMPs, and indeed, proteases belonging to other families are often parts of integral biological and physiological processes, and the inhibitors are often have a broad inhibitory spectrum, much more research is needed to understand the specific functions and mechanisms of proteases as well as the effect of the addition of both endogenous and synthetic inhibitor as a potential therapeutic to target proteases and treat IBD. Additionally, recent research as part of the IPODD consortium (Intestinal Proteases: Opportunities and Drug Discovery) have found that the origin of MMP activity in the gut could actually be attributed to bacteria and also that specific metalloproteases such as Fragilysin from *Bacteroides fragilis,* an organisms generally considered a commensal of the gut microbiota but can also be opportunistically pathogenic, can exhibit similar MMP activity as the host and can significantly contribute to chronic inflammation [103, 104] .

It is not just pathogenic proteases that can be implicated in disorders of the gut. Proteases are also produced by the commensal gut microbiota and there are theories and evidence from studies of pure cultures of gut commensals that suggest they can contribute to damage to the gut [105, 106]. An increased repertoire of proteases provides bacteria with a metabolic advantage which, although it remains to be proven whether

microbial dysbiosis is a cause or effect in IBD, may provide further explanation for how dysbiosis occurs; diet, lifestyle, host genetics etc. may select for bacteria with an enhanced proteolytic repertoire thus are able to out-compete members of the 'normal' gut microbiota that are unable to utilise both exogenous (dietary) and endogenous protein as successfully. Endogenous protein is provided in the form of extracellular matrix protein, mucin, keratin [107], collagen, immune system proteins including immunoglobulins, defensins, cellular adhesion molecules (e.g. E-cadherin) and transmembrane proteins (occludins, claudins, zonula occludins) and protease activated receptors (PARs) [105, 106].

Compromised host epithelial barrier function is a proposed mechanism for aetiology of IBD which has been mentioned previously. The main proteins involved in epithelial cell structure and function that are putative targets of microbial proteases include E-cadherin; a transmembrane adherence protein that joins epithelial cells together, and other protein complexes involved in tight junctions including the transmembrane protein mentioned earlier.
Steck and colleagues have demonstrated that the extracellular metalloprotease secreted by *Enterococcus faecalis*; GelatinaseE (*gelE*) can impair intestinal barrier function in Il-10 deficient mice (Il-$10^{-/-}$). It was also found that *gelE* was capable of degrading recombinant mouse E-cadherin suggesting a defined mechanism for *E. faecalis* involvement in deleteriously altering epithelial barrier function [89]. Other bacteria have also been shown to compromise epithelial barrier function by utilising proteases some of which are shown in Table 1. Therefore there is clear evidence to implicate proteolytic activity in compromising epithelial barrier function which may have implications for IBD and perhaps other gastrointestinal disorders [106]. Also, an area of research that remains relatively overlooked is the level of microbial translocation and associated pro-inflammatory products due to intestinal permeability and contribution to systemic infections in people with disorders of the gastrointestinal tract. Microbial proteolytic activity may be a direct contributor to increased susceptibility to microbial translocation from the gut.

Bacterial proteases may also exert activity on human host cells by way of protease activated receptors (PARs). PARs come in four isoforms (PAR1-4) and are a family of G protein-coupled receptors that are expressed on enterocytes (as well as numerous

other cells including immune cells, neurons and endothelial cells). They are activated (normally) by endogenous serine proteases, in response to a particular stimuli, including inflammation [105] and can contribute to the pro-inflammatory response. Many bacterial species have been shown to be capable of activating host PARs with detrimental effects [108]. The cysteine protease of the oral pathogen *Porphyromonas ginigvalis*; Gingipain has been demonstrated to cleave and activate PAR-1 leading to an up-regulation of pro-inflammatory cytokines[94] and can also activate PAR2 on human neutrophils [106]. PAR2 is overexpressed in mast cells in ulcerative colitis [109] and organisms such as *Pseudomonas aeruginosa* and *Serratia marcescens* secrete proteases (exoprotease LepA and serralysin respectively) that can cleave PAR-1,-2 and -4 and consequently activate NF-κB [107, 110]. Clearly there is a role for dysregulated PAR cleavage and activation in IBD, however it has been demonstrated that host serine proteases are also increased in IBD [111] so more work is to be done to determine the extent of microbial involvement in PAR activation compared to endogenous proteolytic activation.

Microbial proteases have been shown to modulate immune system proteins including immunoglubulins (Ig) (namely secretory IgA, the most abundant Ig in the human gut), cytokines and Toll-like receptors (TLRs) the effects of which have been shown to either enhance the immune response or inhibit it, the consequence of which for both are significant in causing a pro-inflammatory response or enabling bacterial establishment respectively.

The ability to degrade IgA has been demonstrated by numerous bacteria including *Pseudomonas aeruginosa* [112] and *Proteus mirabilis* [113] suggesting a mechanism for evasion of this protective protein. The ability of gut commensals or pathogens to do this would significantly improve their ability to colonise the gut and/or establish an infection.

Evidence from studies of other bacteria that are not considered normal commensals of the gut microbiota suggests a role for disruption of cytokine signalling [105]. For example the trypsin-like cysteine proteases; Gingipans, produced by *Porphyromonas ginigvalis* have been shown to degrade Il-1β, Il-6 and TNFα and can also truncate Il-8 resulting in enhanced neutrophil recruitment exerting a strong pro-inflammatory response at sites of periodontis without elimination of the bacterial infection [114]. Another cysteine protease produced by *Streptococcus pyogenes* (SPE B) can cleave Il-1β in its inactive form to produce active Il-1β in a way analogous to that of endogenous

human cysteine protease consequently increasing Il-10 levels and the inflammatory response [115]. Table 1 includes other bacterial species that target cytokine and other immune system proteins via proteolytic degradation. Clearly there is much evidence that bacteria can utilise their proteases to modulate the host immune response to both enhance or inhibit it, both of which have a deleterious effect on the host and can be implicated in the pathogenesis of IBD.

Mucins, highly glycosylated large proteins that line the gut epithelia may also be considered a protein involved in immunity due to their protective role as a physical barrier in the gut. In IBD, the mucus layer is thinner and has an increased abundance of luminal bacteria associated with it [116]. Consequently, it has been hypothesised that the diminished mucus layer is as a result of degradation by the gut microbiota of which the ability to do so, enabled by both proteases and glycosidases, is also associated with enhanced adherence and disruption of epithelial barrier tight junctions [106, 116]. It is yet to be determined whether the diminished mucus barrier is due to bacterial degradation, or as a result of the disease thus allowing more bacteria to colonise. Further research is in need, but this is still a proposed mechanism for how bacteria with enhanced proteolytic capacity may gain access to the gut epithelium in the first instance and may go on to disrupt barrier function.

The metabolism of exogenous protein in the form of dietary protein can also be implicated in disorders of the gut particularly IBD and colon cancer by the release of toxic metabolic products such as ammonia, amines, sulphur metabolites and phenolic compounds [117]. It has been hypothesised that the products of metabolism of dietary protein by host but more so by microbial proteases, is contributing to an increased pool of these harmful metabolites which thus far have been limited to their implication in colon cancer [117]. While associations between diet and IBD have been studied, the molecular consequences of dietary protein metabolism and its effect on the pool of harmful metabolites have not been studied with regards to IBD.

Previous research has shown that there is elevated microbial serine proteolytic activity in the gut of those suffering from IBS with diarrhoea, but most activity was of host origin [111] [118] and matrix metalloprotease activity is also increased [97] but again it has not been determined whether or not this is host or bacterial in origin. The European

research consortium IBDase has investigated the role of proteases and protease inhibitors in IBD [86] mainly of human origin, but researchers belonging to the 'IPODD' Consortium have found that the activity of many of the proteases, particularly metalloproteases can be attributed to gut bacteria (http://www.efcca.org/index.php/our-activities/latest-news/33-targeting-the-demolition-squad) .

While proteases produced by pathogens are well documented, the extent and capacity of proteolytic activity of commensals, particularly of the gut microbiota has not been comprehensively researched. It is therefore important that the proteolytic capacity of the gut microbiota be determined if we are to examine the impact proteolytic activity is having on the host.

AIMS

The aim of this research is to expand our understanding of proteolytic function within the gut microbiota by comparing proteolytic capacity in faecal samples as a representative of the gut microbiota in healthy individuals with that of IBD sufferers and also characterise the gut proteolytic microbiota in IBD in a relatively large cohort-based study. It is hypothesised that the gut microbiota of IBD sufferers harbours an altered and aggressive commensal bacteria with an increased repertoire of proteases with an increased substrate range. This research will help determine whether proteolysis carried out by the gut microbiota may be a virulence factor contributing to inflammatory disorders of the gut.[119].Additionally, this research will also implement novel tools for an extensive assessment of protease activity in faecal protein extracts which, to the authors knowledge, has not been done before on mixed faecal protein samples nor in an IBD cohort. With the conclusions generated from the research undertaken for Chapter 6.0, the optimal procedure for processing and storing the faecal samples with the overriding purpose of analysing protease activity could also be implemented for this research. Faecal samples could also be stored for extended periods of time at sub-zero temperatures

The main aims of this research are: -

1. To conduct a comprehensive comparison of faecal protease (FP) activities between a cohort of patients with IBD versus healthy controls, and in doing so incorporating exploration of novel tools for assessing total protease activity in faecal samples.

2. To determine which types of proteases are most abundant in each cohort and establish any differences and the origin of FP activity.

3. To begin to unravel the potential that FP activity may be implicated as a virulence factor in IBD by analysis of their effect on human cell lines.

4. To conduct a comparison of gut microbial communities between each cohort using 16S rRNA gene sequencing analysis and to determine any correlations between certain bacterial taxonomies with elevated or low FP activity.

7.2 RESULTS

7.2.1 Protein concentration in healthy and IBD protein extracts

There was no significant difference found between the total average of protein concentrations of all the healthy samples and all the IBD samples (Figure 7.2.1).

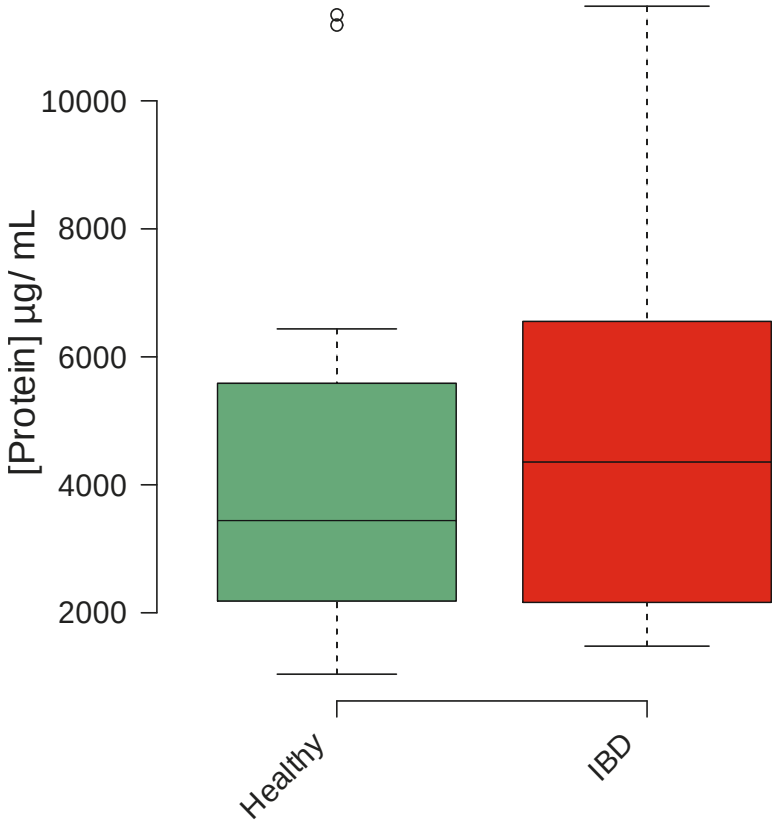


Figure 7.2.1 Average protein concentrations after extraction of the total cohort of healthy compared with the IBD patients. Center lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by dots; width of the boxes is proportional to the square root of the sample size.

Healthy(n) = 24, IBD(n)= 26 sample points which include total protein and extracellular protein samples**.**

7.2.2 Feacal Protease activity in Patients with IBD compared to Healthy controls

General protease activity was assayed colourimetrically using azo-dye impregnated casein as the substrate. The IBD samples showed significantly higher caseinolytic activity (Figure 7.2.2). Keratin azure (azure dye-impregnated sheep's wool keratin) was used to determine keratinolytic activity. The IDB samples showed significantly higher keratinolytic activity when compared to the healthy samples.

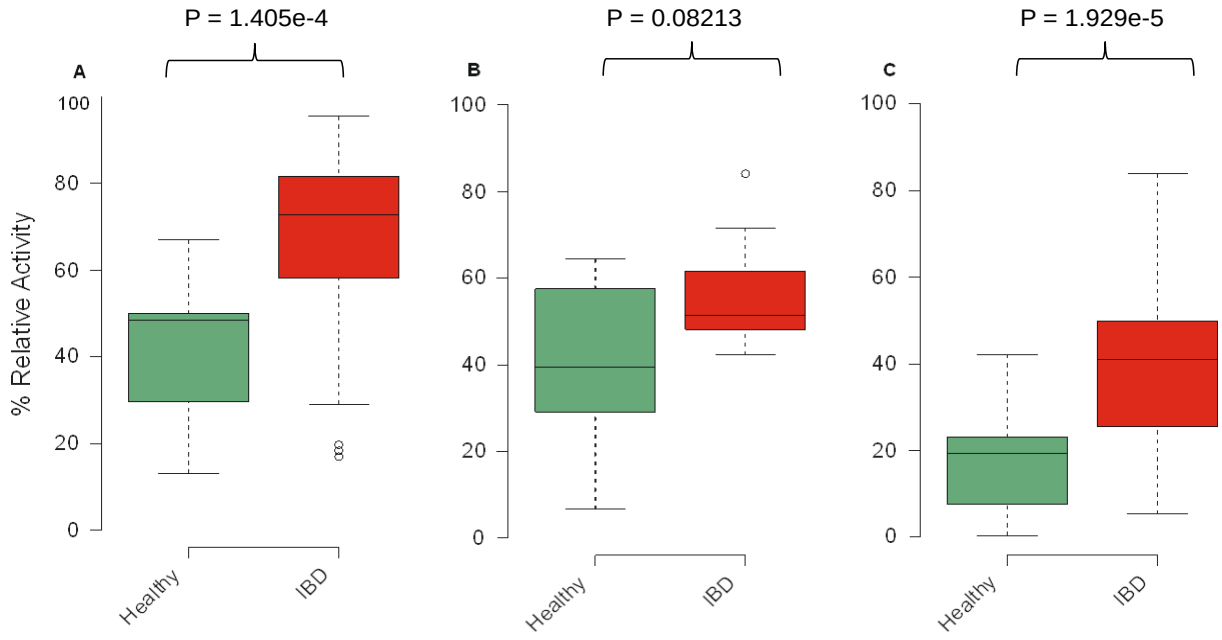
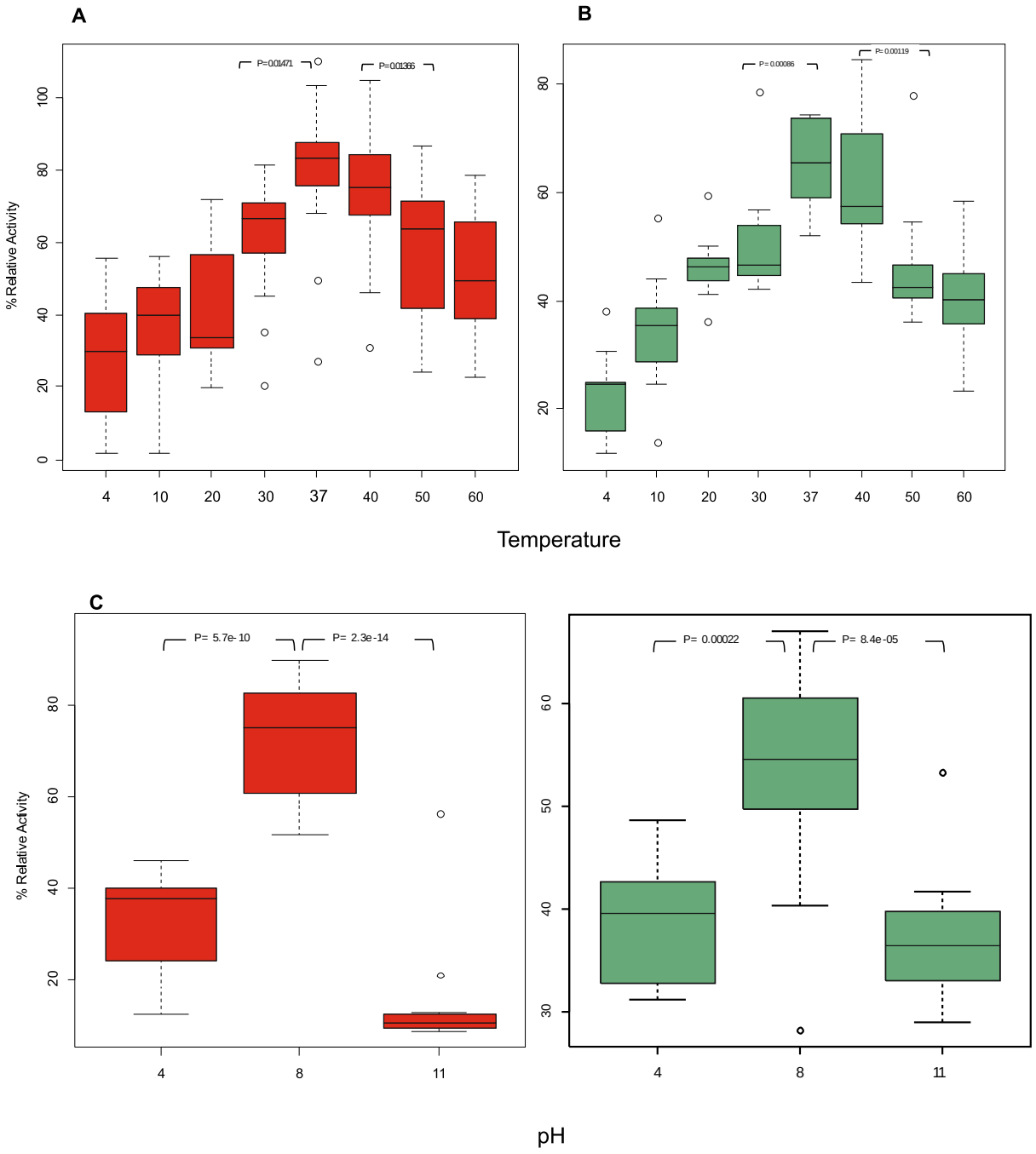


Figure 7.2.2. **A**=Azo-casein assay. **B**= Azo-coll assay, **C**= Keratin azure assay. Results represent the mean of triplicate data per sample and then the mean of the entire cohort of healthy samples (**green** boxplot on the left of each plot) compared with the entire cohort of IBD samples (**red b**oxplot on the right of each plot). Statistical significance between FP activities in Healthy compared to IBD was determined using the Mann-whitney U test in R software.

7.2.3 Temperature and Ph profiles of FP activity

FP activity was significantly higher at 37 °C for each cohort (figure 7.2.3 **A** & **B**) compared to the activity at all other temperatures apart from activity at 40 °C which was also signifcantly higher than activity at all other temperatures.Optimum FP activity was observed at pH 8.0 (at a temperature of 37 °C) for both cohorts (Figures 7.2.3 **C** & **D**).



←*from previous page*

Figure 7.2.3. Temperature (A & B) assays and pH (C & D) assays to determine the optimum temperature and pH respectively for FP activity Results represent the mean of triplicate data per sample and then the mean of the entire cohort of healthy samples (**green** boxplot on the right of each figure) compared with the entire cohort of IBD samples (**red** boxplot on the left of each figure). Statistical significance between FP activities at the various temperatures and pH were determined using ANOVA and continuous pairwise comparisons in R software.

7.2.3 Protease Inhibitor assays

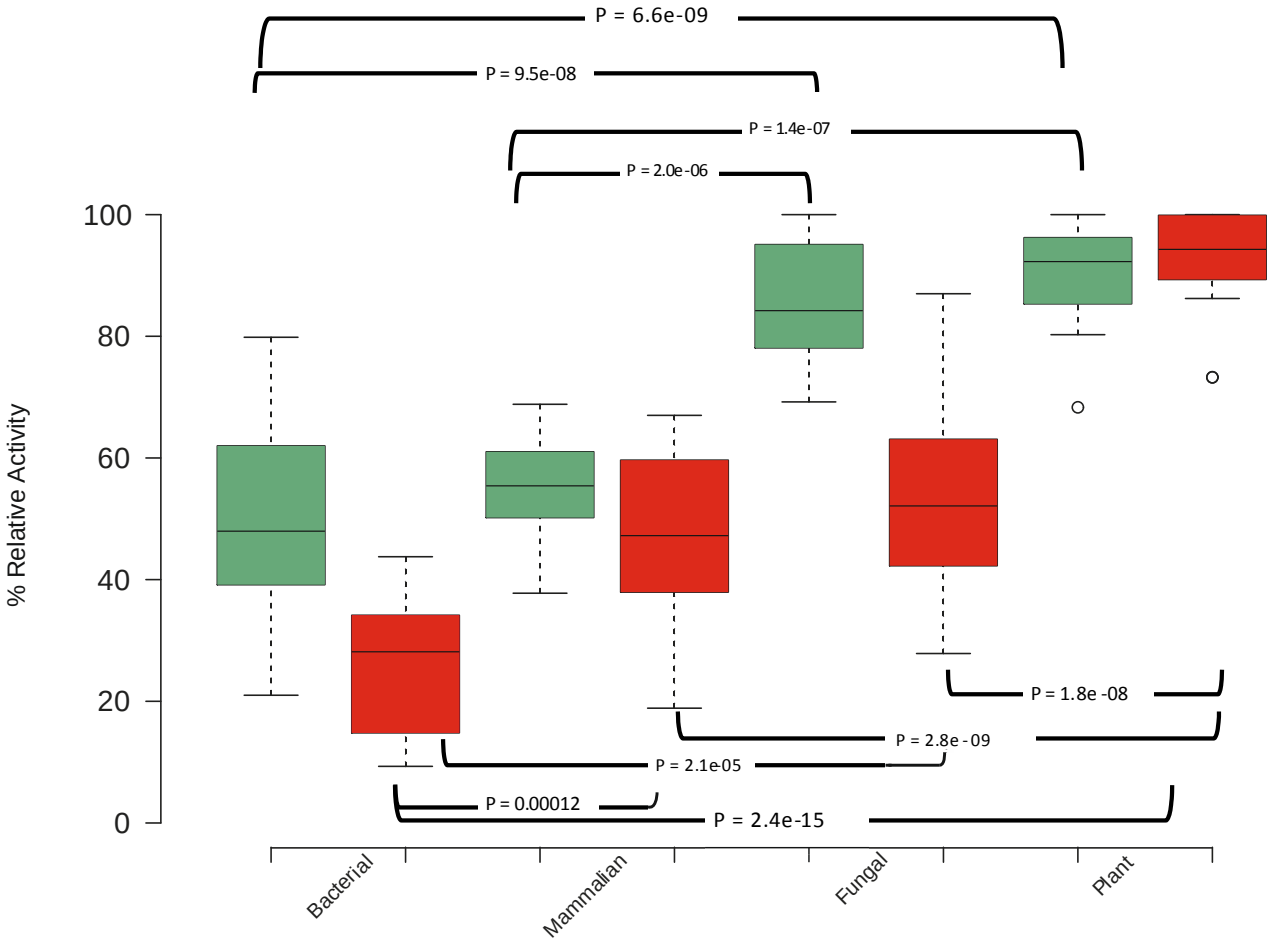
ProteaseARREST™ (G-Biosciences) species specific inhibitor cocktails were used to test the samples for presence of bacterial, mammalian, fungal and plant proteases by assessing relative activity after incubation with the inhibitor. A lower activity was representation of a greater level of inhibition indicating greater levels of proteases belonging to that species. All samples (healthy and IBD) showed greatest inhibition of activity upon addition of the bacterial protease inhibition cocktail (figure 7.2.4) and in the presence of mammalian proteases inhibitors, activity was also diminished in all samples (figure 7.2.4). In the presence of fungal protease inhibitors, the healthy samples maintained high levels of activity however they had a negative effect (less than 50%) on over half of both the IBD types (table 7.1). None of the samples were negatively affected by the presence of the plant specific protease inhibitor cocktail (figure 7.2.5 and table 7.1).

As it appeared that bacterial proteases were responsible for a significant level of FP activity from this assay, to determine the contribution on intracellular bacterial proteases to the observed activities an Optiprep™ density gradient medium was used to isolate bacterial cells and then determine FP activity before and after cellular disruption by bead beating. It was found that following cellular disruption significantly higher % relative activity was observed (figure 7.2.4B, $P<0.05$) suggesting that intracellular bacterial proteases were contributing to the FP levels observed from previous assays but disrupted bacterial cells exhibited significantly lower FP activity compared to total protein extract and extracellular protein extract ($P<0.05$) indicating that secreted proteases are responsibly for a significant level of the activity observed. Total protein and

extracellular protein were not significantly different suggesting that after normalisation of protein, the samples were composed of a similar abundance of proteases.

To determine which proteases families were most abundant in each sample and whether there were differences between healthy and IBD, a series of protease inhibitors were used to assess activity after incubation with the following inhibitors; AEBSF, ALLN, Antipain-dihydrocholride, Aproptinin, Bestatin, Chymostatin, E-64, EDTA, Leupeptin, Pepstatin, Phosphoramidon and PMSF. Each of these inhibitors are specific for a particular family of proteases for example PMSF is a serine protease inhibitor and phosphoramidon inhibits metalloproteases. The serine protease inhibitors AEBSF, Aprotinin, chymotrypsin inhibitor, PMSF and the metal ion chelator EDTA-Na caused significant inhibition of activity compared to FP activity when no inhibitor was present in the healthy cohort (Figure 7.2.5). Similarly, these inhibitors were also responsible for a significant reduction in FP activity in the IBD cohort too, apart from in the presence of chymotrypsin-like inhibitor (figure 7.2.5) however, in addition to these, the trypsin-like serine protease and cysteine protease inhibitor Antipain dihydrochloride, and the metalloprotease inhibitor phosphoramidon, also caused significant reduction to FP activity. In the presence of phosphoramidon, IBD FP activity was significantly different when compared to the healthy cohort ($P=0.0097$) suggesting a greater level of inhibition of FP activity due to phosphoramidon in the IBD cohort compared to the healthy controls.

A


P = 6.6e-09
P = 9.5e-08
P = 1.4e-07
P = 2.0e-06
100
80
60
40
20
0
% Relative Activity
P = 1.8e -08
P = 2.8e -09
P = 2.1e-05
P = 0.00012
P = 2.4e-15
Bacterial
Mammalian
Fungal
Plant

B

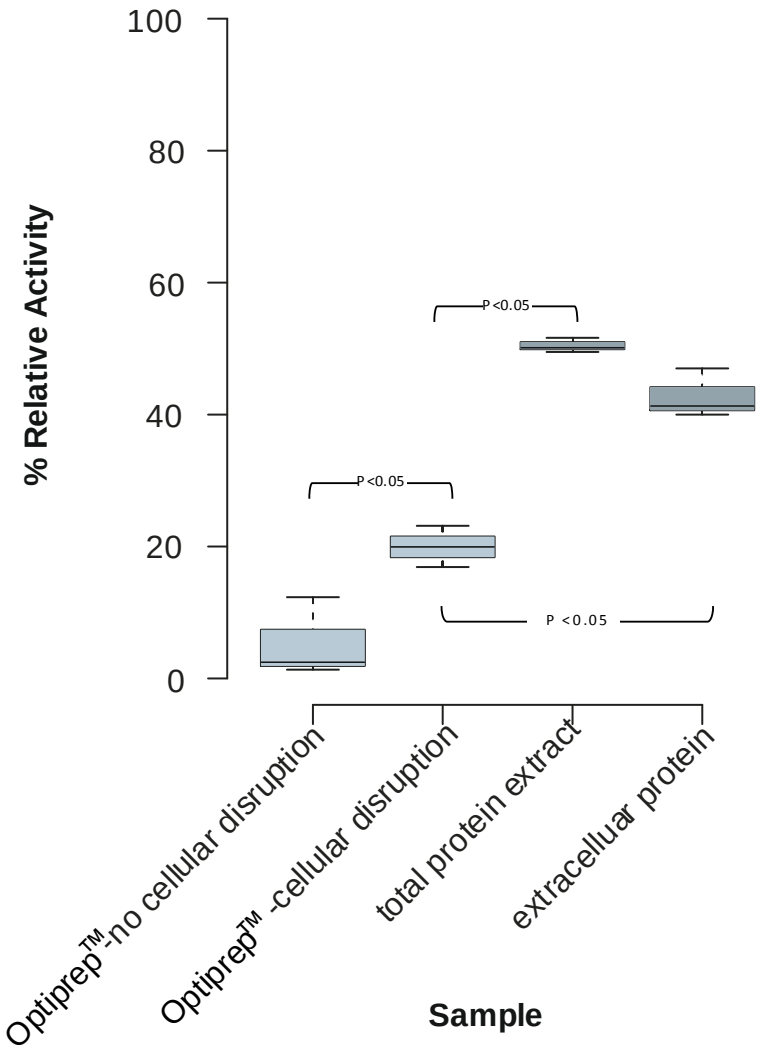
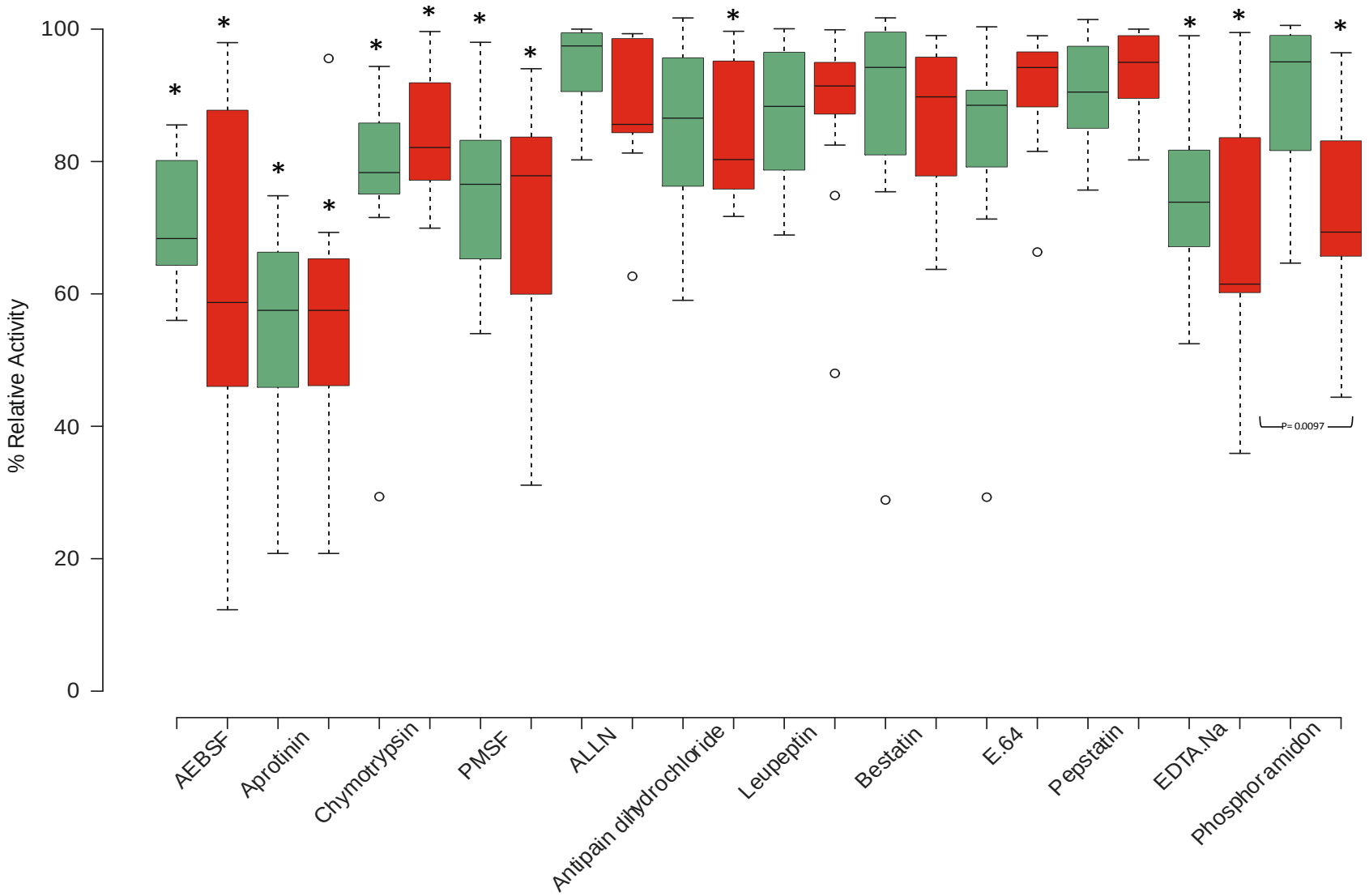


Figure 7.2.4**A**: ProteaseARREST™ assays for Bacterial, mammalian fungal and plant specific protease inhibition (from left to right). Results represent the mean of triplicate data per sample and then the mean of the entire cohort of healthy samples (green boxplots) compared with the entire cohort of IBD samples (red boxplots. Centre lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by dots.N = 11, 13, 11, 13, 11, 13, 11, 13 sample points. High activity is indicative of less inhibition due to the presence of the protease inhibitor cocktail; conversely, low activity is indicative of inhibition by the inhibitor and indicates the presences of that type of protease in the sample. Statistical significance between FP activities in Healthy compared to IBD was determined using the Mann-

whitney U test in R software. Differences between activity within each cohort in the presence of the different inhibitors was assessed by ANOVA and pairwise comparisons. **B:** Relative proteolytic activity (%) following isolation of bacterial cells using an Optiprep™ density gradient method with and without cellular disruption. Relative proteolytic activity was compared with that of total protein protease activity and extracellular protein protease activity without bacterial cell isolation. Bacterial specific protease activity was significantly higher after cellular disruption ($P<0.05$). However, protease activity was significantly higher in the total protein and extracellular protein extracts ($P<0.05$).


% Relative Activity
100
80
60
40
20
0
P= 0.0097
AEBSF
Aprotinin
Chymotrypsin
PMSF
ALLN
Antipain dihydrochloride
Leupeptin
Bestatin
E.64
Pepstatin
EDTA.Na
Phosphoramidon

←*From previous page*

Figure 7.2.5. FP activity from the healthy cohort (green boxplots) and the IBD cohort (red boxplots) after incubation with different protease inhibitors (at 37°C and a neutral pH). Centre lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by dots. n = 11, 13, 11, 13, 11, 13, 11, 13, 11, 13, 11, 13, 11, 13, 11, 13, 11, 13, 11, 13, 11, 13, 11, 13, 11, 13 sample points.
% Relative activity was calculated compared to the no inhibitor present controls. Lower points are indicative of low activity and therefore higher levels of inhibition meaning that there are proteases from the family of proteases targeted by that inhibitor in that particular sample. Asterisk (*) indicates FP activity that is significantly different from no inhibitor controls: Healthy:- AEBSF; P = 0.00102, Aprotinin; P = 0.00041, Chymotrypsin inhibitor; P = 0.01509, EDTA-Na; P = 0.03015 and PMSF; P = 0.01509. IBD:- AEBSF; P = 0.00344, Aprotinin; P = 7.7e-05, Antipain dihydrochloride; P = 0.04532, EDTA-Na; P = 0.00291, Phosphoramidon; P = 0.00174 and PMSF; P = 0.00174.

7.2.4 Specific protease activity profiles

To futher assess the substrate specificity of each cohort, a set of three fluorometric substrate were chosen following results from the inhibitor study. Mca-Lys-Pro-Leu-Gly-Leu-Dpa-Ala-Arg-NH2 (MMP) was used as a substrate for matrix metalloprotease activity and showed significantly elevated activity in the IBD cohort (figure 7.2.6 A). MeOSuc-AAPV-AMC (MEO) is a substrate for neutrophil elastase activity for which was signficnatly greater in the IBD cohort once again (figure 7.2.6 B). Suc-Ala-Ala-Pro-Phe-AMC was used as a synthetic substrate specific for chymotrypsin activity. No differences were observed between the two cohorts (figure 7.2.6 C).

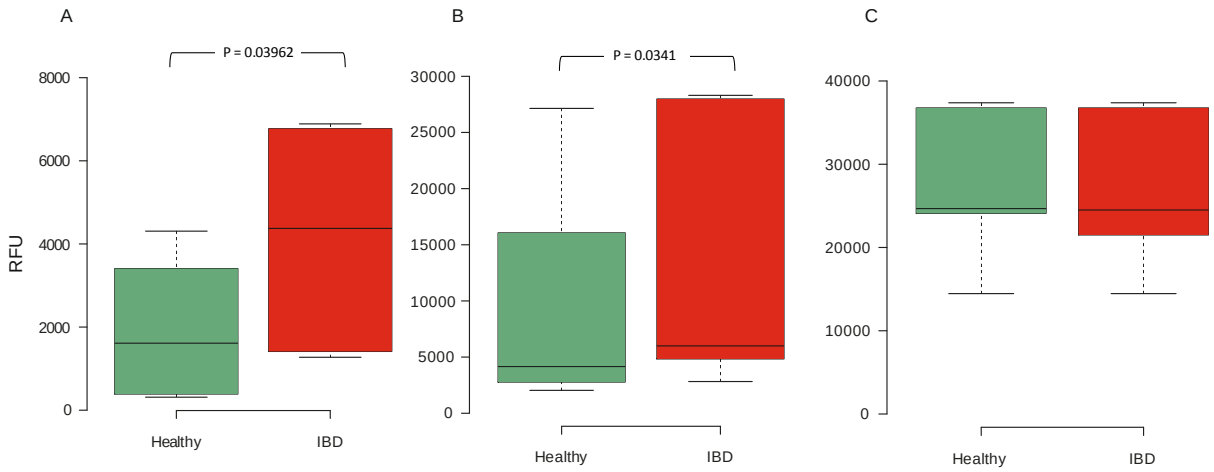


Figure 7.2.6. Relative Fluorescent Units (RFU) released as a result of FP activity from the healthy cohort (**green** boxplots) and the IBD cohort (**red** boxplots) after incubation with fluorescent labelled substrates **A**; MMP, **B**; MEO and **C**; Chymotrypsin. Center lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by dots. n = 11, 13 sample points. Significant differences between the activity of the two cohorts are shown (MMP; P = 0.03962 and MEO; P = 0.0341). P values with no statistical significant difference are not shown.

7.2.5 Degradation of mucin

In order to determine the ability of the samples to degrade mucin, firstly mucin was labelled using biotin through the protein moiety as described in materials and methods. Samples were run through a Sephadex G25 column following biotinylation and a slot blot assay with Periodic acid Sschiff's stain was perforemd to test the collected fractions for mucin and labelling was checked by adhering the mucin to a 96-well plate and streptavidin labelling. Using samples that were successfully biotin labelled, a mucinase assay was performed using the protein extracts from each sample in the two cohorts. No statistically signficnat difference was observed in mucinase acitivity between the two cohorts.

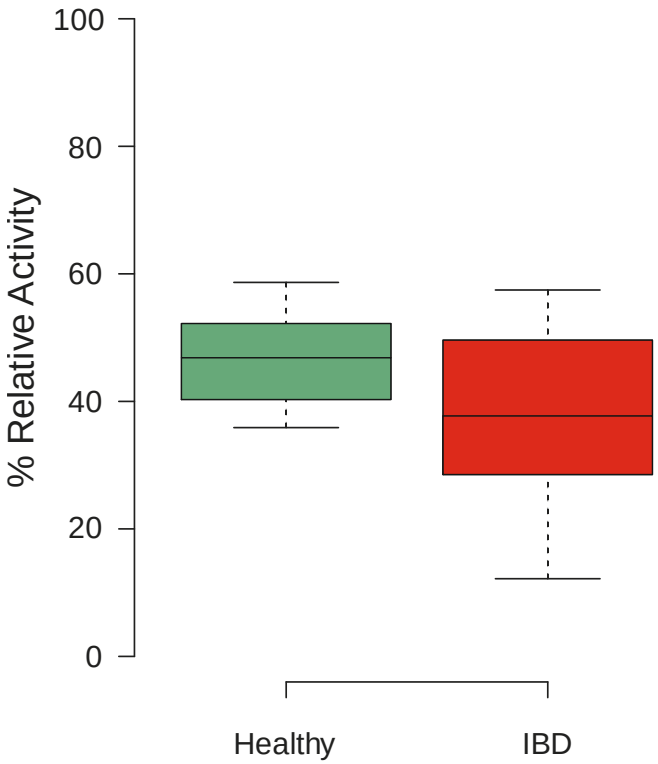


Figure 7.2.7. Relative activity of mucinase in the healthy cohort compared with the IBD cohort to trypsin control. Centre lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, n = 11, 13 sample points. There was no significant difference between the relative activities of each cohort.

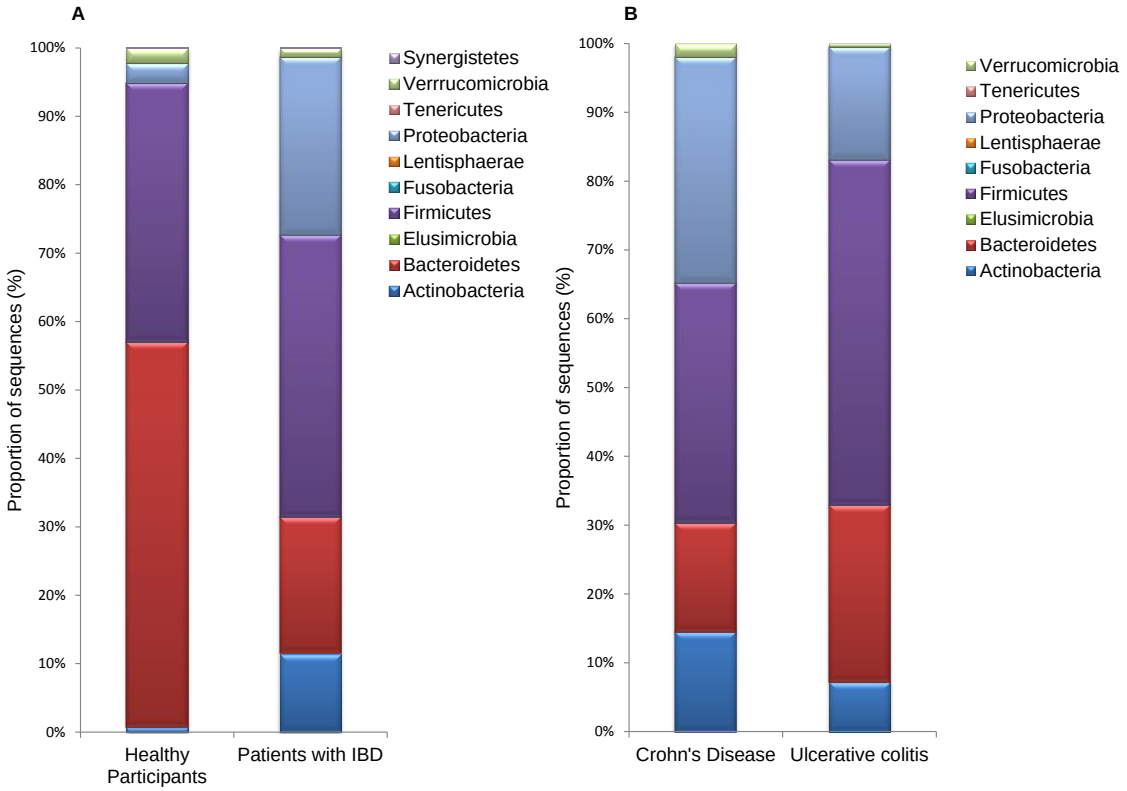
7.2.6 Bacterial Community in the Healthy cohort and IBD patient cohort

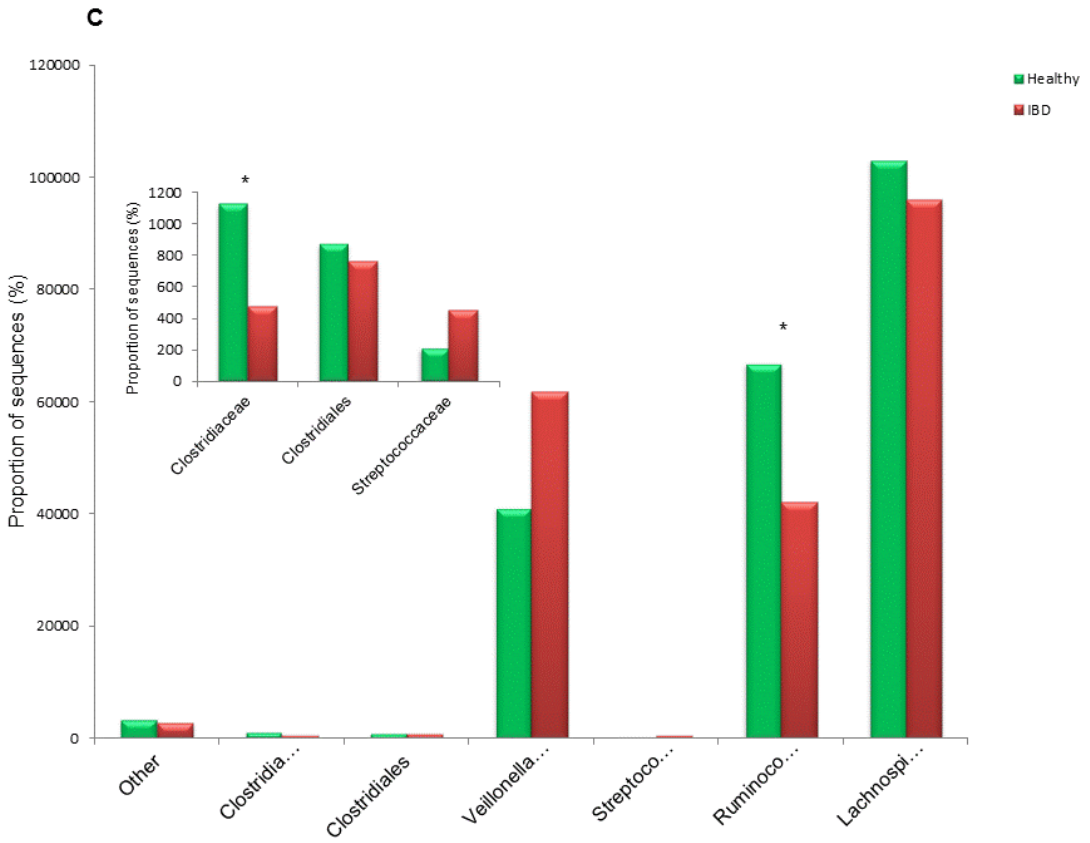
A total of 1426461 high-quality sequences were produced in this study using Illumina Miseq™ DNA sequencing technology with an average of 59436 sequences per sample. The Shannon diversity index and community richness Chao index were calculated for each cohort and are shown in table 7.2. Analysis of the Shannon index for diversity revealed a higher diversity in healthy faecal samples compared to IBD faecal samples, however, this was not statistically significant. When separated into their respective disease phenotypes (CD or UC) samples showed a similar diversity indices and species richness.

The microbial structure of the two cohorts and the IBD cohort separated in respective disease type (CD or UC) to the phylum level are shown in figure 7.2.7A. The dominant phyla in the healthy groups were *Firmicutes, Bacteroidetes* and more prominent in the IBD sample groups were *Proteobacteria* and *Actinobacteria* were also dominant in the IBD sample group. However no statistically significant differences were observed at the phylum level. There were significant differences between healthy and IBD faecal samples at different bacterial taxonomic ranks namely class and family rank (table 7.3). The microbial structure of the two different IBD types (CD and UC) used in this study are shown in figure 7.2.7B. The dominant phyla in both groups were *Proteobacteria, Firmicutes, Bacteroidetes* and *Actinobacteria. Bacteroidetes* appeared reduced in the CD group compared to the UC group and conversely, *Actinobacteria* were increased in the CD group and decreased in the UC group.

Table 7.3. 16S r DNA sequencing data summary comparing the IBD patients faecal microbial diversity with the healthy participant faecal microbial diversity. Shannon index form diversity was calculated for each group and was found to be statistically significantly different (P= 0.00?). The Chao index was used to determine the richness of OTUs in the dataset.

	Healthy (n = 11)	IBD (n = 13)	CD (n = 7)	UC (n = 6)
# Sequences	50794±3069	66747±51124	78459±18194	53084±36585
# Filtered OTUs	475±75	464±98	410±152	397±164
Chao	524±205	395±241	404±174	384±230
Shannon	2.627±2.755	2.295426±3.01	2.275±2.766	2.319±2.567


A
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Proportion of sequences (%)
Healthy Participants
Patients with IBD
Synergistetes
Verrrucomicrobia
Tenericutes
Proteobacteria
Lentisphaerae
Fusobacteria
Firmicutes
Elusimicrobia
Bacteroidetes
Actinobacteria
B
Crohn's Disease
Ulcerative colitis
Verrucomicrobia
Tenericutes
Proteobacteria
Lentisphaerae
Fusobacteria
Firmicutes
Elusimicrobia
Bacteroidetes
Actinobacteria


C
120000
100000
80000
60000
40000
20000
0
Proportion of sequences (%)
Healthy
IBD
1200
1000
800
600
400
200
0
Proportion of sequences (%)
*
Clostridiaceae
Clostridiales
Streptococcaceae
*
Other
Clostridia…
Clostridiales
Veillonella…
Streptoco…
Ruminoco…
Lachnospi…

← *From previous page*

Figure 7.2.8. A; % Relative abundance of the main bacterial phyla after analysis of 16S Ribosomal DNA gene sequences from faecal samples of healthy participants and patients with IBD. The IBD cohort is characterised by a reduction in *Bacteroidetes* and an expansion of *Proteobacteria* and *Actinobacteria*. **B**; % Relative abundance of the main bacterial phyla after analysis of 16S Ribosomal DNA gene sequences from faecal samples from the IBD cohort separated into their respective disease types (CD and UC). *Firmicutes, Proteobacteria, Actinobacteria* and *Bacteroidetes* are again, the dominant phyla. % Proportion on *Actinobacteria* is higher in CD than UC, similarly are *Proteobacteria*. **C**; Family level comparison of *Firmicutes* in the two cohorts. *Clostridiales* encompasses *Clostridiales insertae sedis* XI,XII and XIII, *Clostridiaceae* encompasses *Clostridiaceae* 1,2 and 4 and the group 'other' encompasses *Aerococcaceae, Carnobacteriacaea, Erysipelotrichaceae, Enterococcaceae, Eubacteriaceae, Gracilibacteriaceae, Heliobacteriaceae* and *Peptostreptococcaceae.* Samples that were found to be significantly different between the groups within the *Firmicutes* Phylum are indicated by an '*'. The *Clostridiaceae*, *Clostridiales* and *Streptococcaceae* are shown in the smaller insert graph with a different scale for proportion of sequences.

Table 7.3 Bacterial groups that are significantly different between the two cohorts as determined using an adjusted P-value as determined by two groups comparisons using white's non-parametric t-test and Benjamin-Hochberg false discovery rate to account for multiple comparisons

Taxonomy	Rank	p value	Adjusted p value	%IBD	% Healthy
Alphaproteobacteria	Class	9.99e-4	8.99e-3	0.100712	2.58415
Bacilli	Class	9.99e-4	0.018	0.592636	0.064676
Gammaproteobacteria	Class	6.99e-3	0.042	12.5288	0.0345518
Actinobacteria	Class	9.99e-3	0.045	11.6796	0.95237
Enterobacteriaceae	Family	2.39e-4	0.011	12.4971	0.00611988

7.2.7 Association of Faecal Microbial Diversity and Taxonomy with protease activity

In order attempt to determine any microbial taxonomic associations with high and low levels of faecal protease activity, the data obtained from general protease activity assays was used to divide the samples from the entire study group into quartiles for high (n = 6) and low activity (n = 6).

The distance matrix that had previously been calculated in Mothur [120] was analysed by nonmetric multidimensional scaling (NMDS) to enable a qualitative visualisation of the similarity between community structure in high FP activity compared to low FP activity. Analysis of the NMDS showed separation in high vs low FP activity when the cohort of high FP was compared to the cohort of low FP activity (figure 7.2.11 A). When the IBD and healthy samples were analysed they too revealed a separation between samples with high FP and low FP activities (figure 7.2.11 B and C) notably in the healthy cohort, the high FP samples were much more similar than the high FP samples that showed much more variation.
Comparison of the microbial communities of the samples with high and low activities was conducted in STAMP [121] and R software [122]. Shannon index for diversity was calculated for each group and was found to be statistically significantly reduced in the high FP group compared to the low FP activity group ($P<0.05$, figure 7.2.10).

STAMP software was used to determine if there were any compositional differences in the faecal microbiota of the high activity compared to low activity sample groups. Compositional differences of associations with high or low FP activity were observed at the Phylum level (figure 7.2.12).

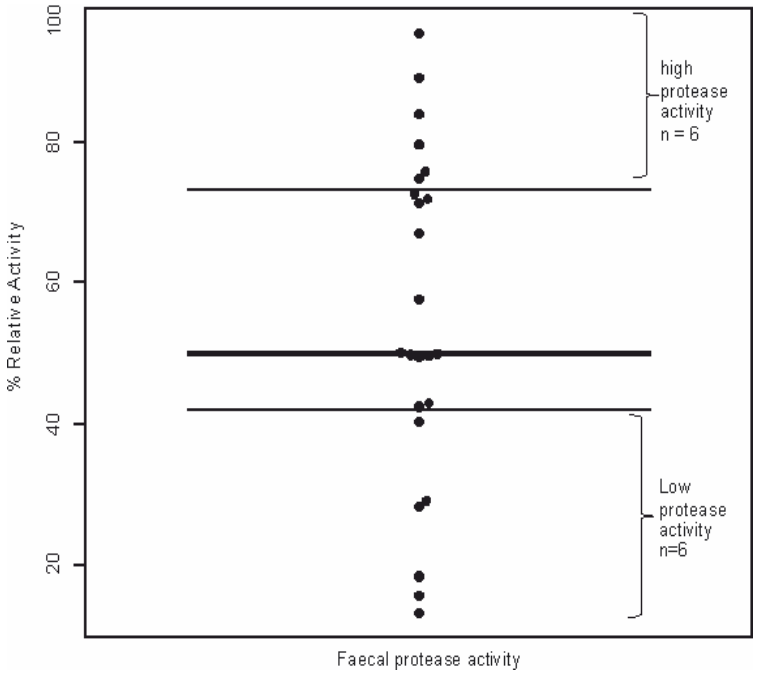


Figure 7.2.9. FP activity in faecal protein extracts from all study participants (not separated into disease type or healthy). The bottom quartile represents sample with low FP activity (n = 6) as determined by R software and the top quartile represents the samples with high FP activity (n = 6). Samples not in either quartile were deemed insignificantly representative of high or low FP activity and so were not included in further analysis.

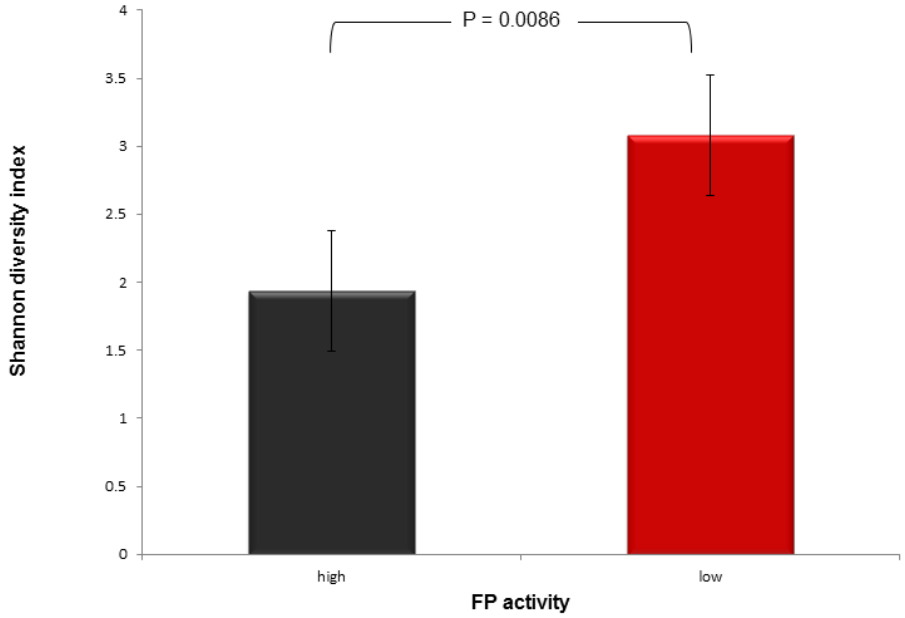


Figure 7.2.10 Microbial diversity of sample with high and low FP.

Shannon index of diversity was calculated for each group and was found to be statistically significantly different (P= 0.00866) between high and low FP groups.

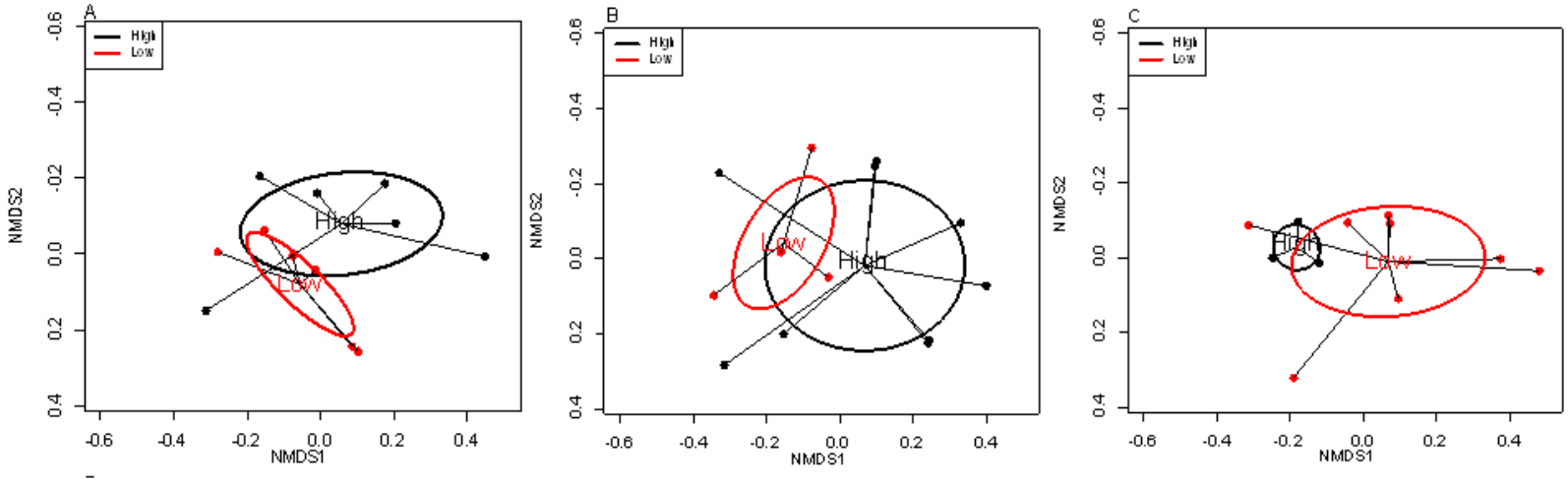


Figure 7.2.11 NMDS map showing the separation of samples that exhibit high FP activity from the samples that exhibit low FP activity. A; the entire cohort (both healthy and IBD after statistical speartion into cohorts of high and low FP activity). B; the IBD cohort separated into high and low FP activity and C; the healthy cohort separated into high and low FP activity.

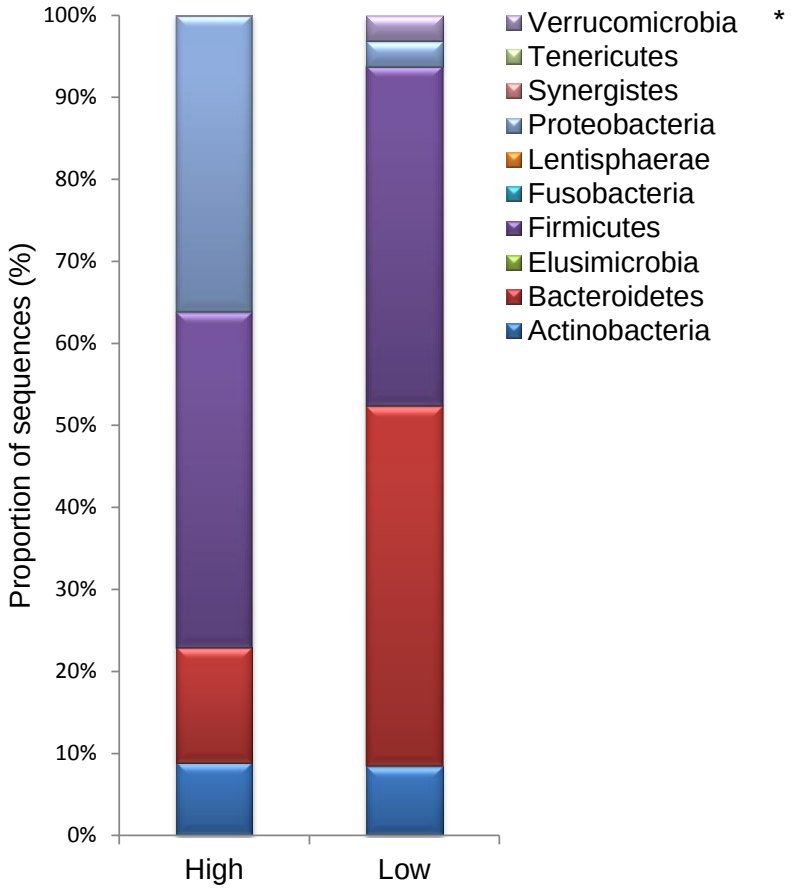
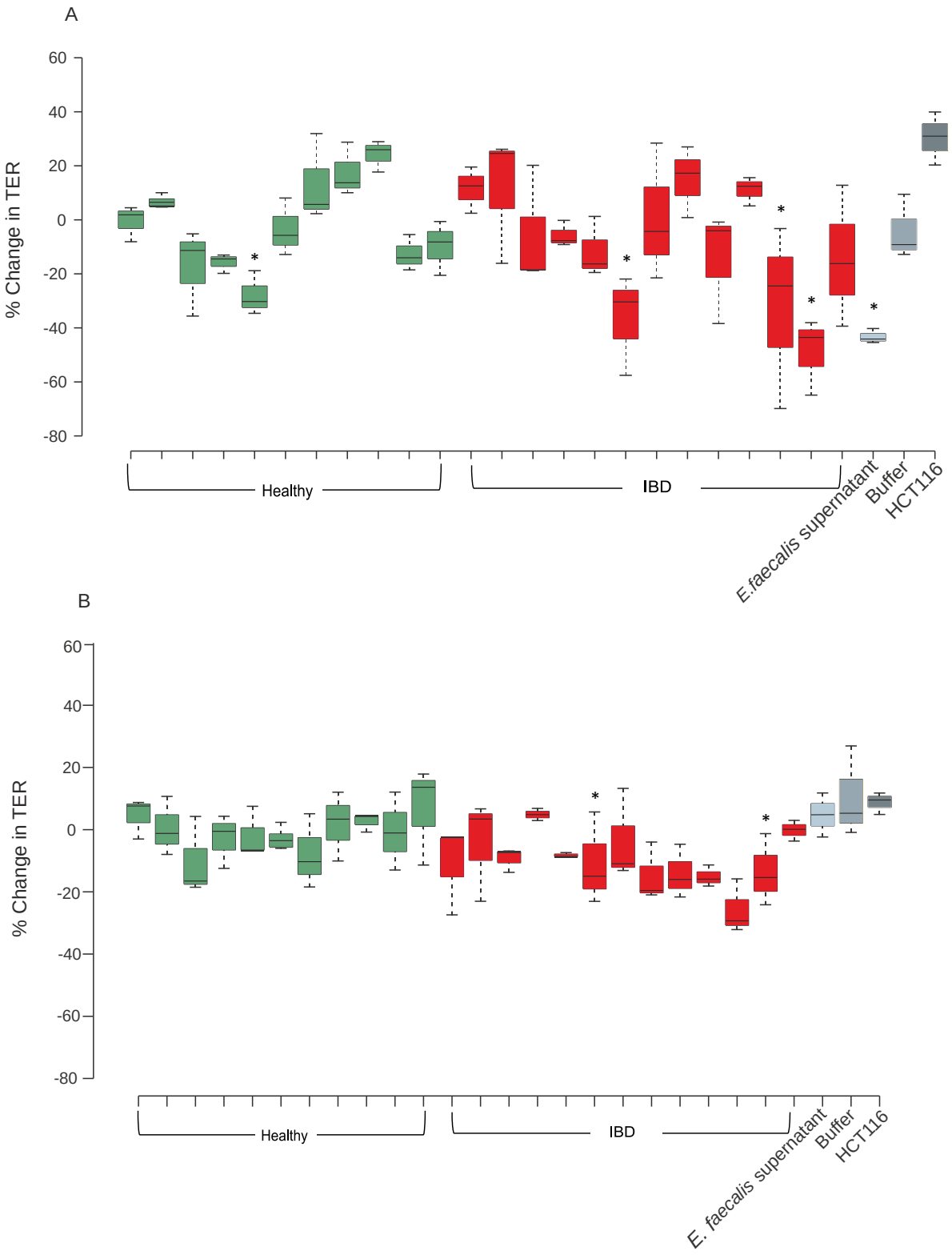


Figure 7.2.12. % Relative abundance of the main bacterial phyla after analysis of 16S Ribosomal DNA gene sequences from faecal samples of healthy participants and patients with IBD separated into quartiles of high and low FP activity. The high FP activity cohort is characterised by a reduction in *Bacteroidetes* and an overall reduction in diversity and an expansion of *Proteobacteria*. The low FP activity cohort can be characterised by the presence of members from other Phyla, and dominance by *Bacteroidetes* and an increase in *Verrucomicrobia*. Phyla found to be significantly different in abundance between the two groups are marked with an '*'.

7.2.8 Effect of protease activity on TER

To begin to unravel the role of the gut microbiota and their proteases as virulence factors, the effect of faecal protease extract from both the cohorts on the epithelial barrier integrity of two human gut epithelial cell lines (HT-29 and HCT116) was investigated by measuring the TEER and the translocation of fluorescent labelled sodium fluorescein molecular markers after 24hours of incubation. Concentrated *E. faecalis* crude protein extract was used as a positive control as this has been previously shown to decrease TER [89]. Protein extract buffer and a sample with of cell culture alone were also used as negative controls. Samples were first compared to determine which showed a statistically significant decrease in TER with the control, these samples were further compared before and after 24hours of incubation to determine if this change was significant. Healthy sample 5 and IBD samples 6, 11 and 12 showed a significant decrease in TER after incubation ($P = 0.02856$, 0.04898, 0.02815 respectively) as did the concentrated *E. faecalis* supernatant ($P = 0.0002566$). IBD samples, 6, 11 and 12 also showed a higher translocation of a sodium fluorescein molecular marker from the apical side to the basolateral side of the cell monolayer when no inhibitor was present compared to when the inhibitor was present (figure 7.2.13; $P=$ 0.0369, 0.00063 and 0.0463 respectively) To confirm whether the decrease in TER was protease mediated, in parallel the Transwell cultures were also incubated with protein extract from each sample and a protease inhibitor cocktail. These samples were subject to the same statistical test to determine if TER was significantly altered after the incubation period. None of the samples showed a significant decrease in activity (figure 7.2.13 B) thus indicating that the cell monolayer was being effected by protease activity and not some other factor. A significant ($P <0.001$, $rho = -0.763$) negative correlation between FP activity and the % of original TER after incubation with protease extract for 24 hours (figure 7.2.13 C).

Healthy sample 5 did show translocation of the fluorescent marker but the sample with and without inhibitor were not statistically significantly different (figure 7.2.14) furthering the possibility that an alternative enzymatic activity was also responsible for the decrease in TER and epithelial permeability. Healthy sample 6 and IBD sample 9 however, also showed significant translocation with no inhibitor compared to when the inhibitor was present.


A
% Change in TER
60
40
20
0
-20
-40
-60
-80
Healthy
IBD
E.faecalis supernatant
Buffer
HCT116
B
% Change in TER
60
40
20
0
-20
-40
-60
-80
Healthy
IBD
E. faecalis supernatant
Buffer
HCT116

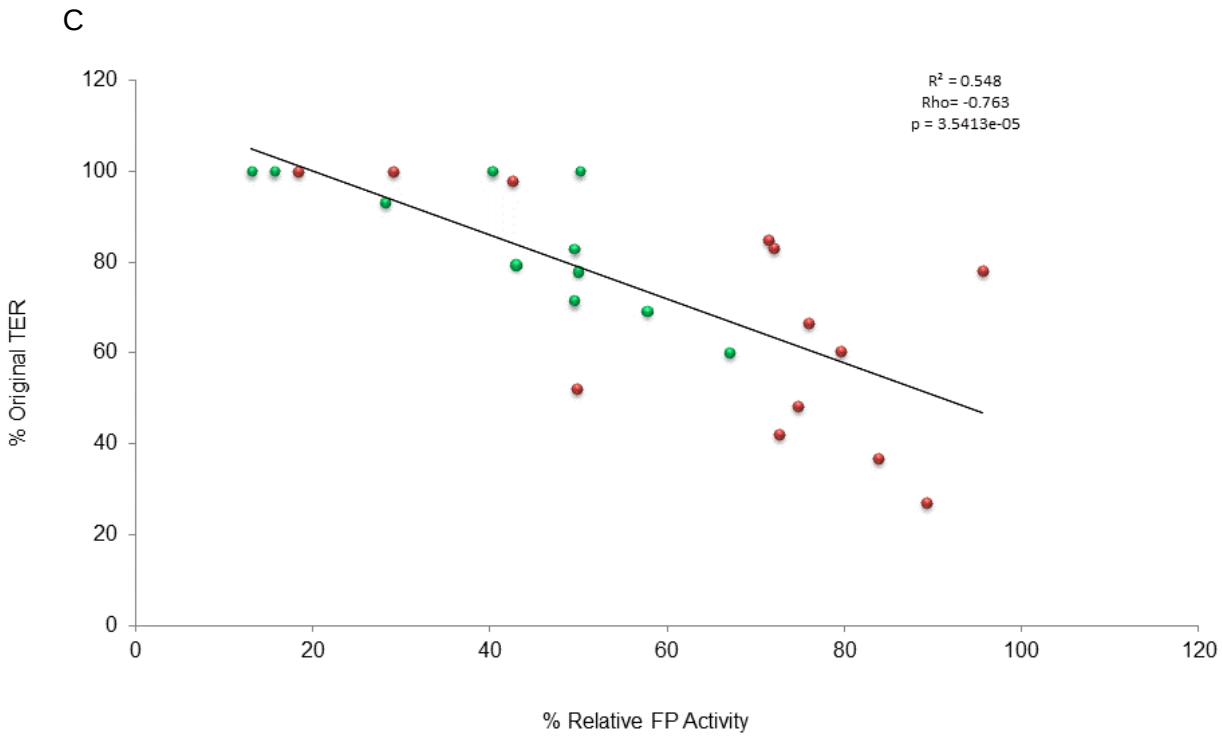


Figure 7.2.13 Effect of FP activities from protein extracts on TER of Ht-29 cell line without (**A**) an inhibitor present and with (**B**) inhibitor after 24 hours of incubation. TER measurements were made at time '0' and after 24 hours incubation and the % change in TER was plotted. Centre lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, n = 11(healthy, which are represented by the **green** circles), 13(IBD, which are represented by the **red** circles) sample points. Samples that had a significant change in TER after incubation compared to their original reading are indicated with an '*'.

As no samples showed a significant decrease in TER in the presence of a potent protease inhibitor, samples were subject to correlation and regression analysis to determine whether FP activity and decrease in TER were associated with one another (**C**). A significant negative correlation was observed (rho = -0.763, P = 3.541e-05).

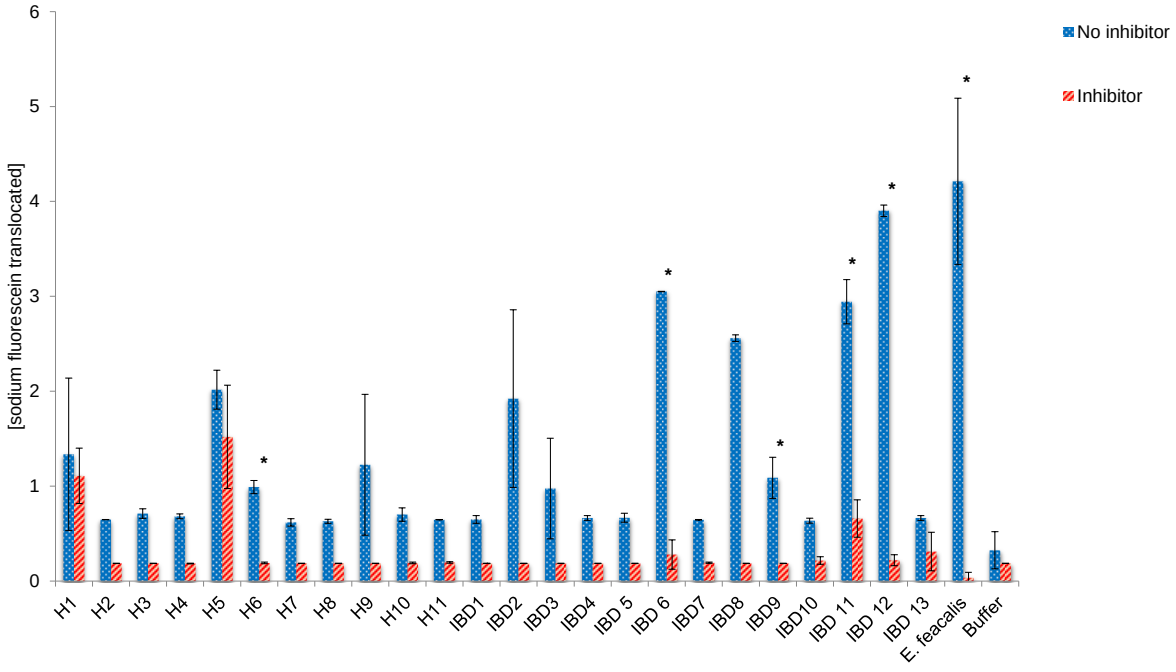


Figure 7.2.14. Permeability assay showing the translocation of sodium fluorescein molecular marker across a monolayer of HCT116 after 24 hours of incubation with protein extract from each sample from both cohorts. Concentrated supernatant of *E. feacalis* was used as a positive control and cells incubated with the protein buffer were used as the negative control and to ensure the buffer was having no effect on cell layer permeability. Samples with '*****' represent samples that had a significantly different concentration of sodium fluorescein translocate across the apical layer to the basolateral layer upon addition of a protease inhibitor cocktail.

7.3 DISCUSSION

The aim of this research was to conduct a novel, comprehensive analysis of the gut degradome, hypothesised to be a virulence factor in patients with IBD; a debilitating, chronically recurring disease of the gastrointestinal tract of prominence in the developed world and associated with early onset (ages 15-30) [1]. Recent years have seen many advances in understanding the aetiology of this disease with regards to genetics, environmental factors, such as diet and smoking and importantly, alterations in and actions of the gut microbiota and the implementation of germ-free or gnotobiotic animal models, however the pathogenesis of the disease remains far from fully being understood. Genome wide association studies have identified susceptibility loci and certain genes that contribute to an individual's predisposition to develop the disease. In broad terms, identified loci comprise genes that have been demonstrated to be of importance in; immune response to bacterial stimuli, autophagy pathways and epithelial barrier function [63]. One of the first genes to be identified in such studies was a frameshift mutation in the *NOD2* gene, a gene involved in immune regulatory pathways such as the activation of nuclear factor (NF)-κB [1, 38]. Since then, many more susceptibility loci and genes have been identified including *JAK2, STAT3, ATG16L1*, and *ECM1*[63]. The implementation of animal models with modified or deleted genes identified to be implicated in IBD susceptibility have helped provide evidence for the implications of dysfunctional genes [123] [124] but one of the key findings from animal models is that, in the absence of a gut microbiota, IBD does not occur. However, what remains unclear is the level of bacterial involvement in the disease; is there a singular infectious microbial agent? (a pathogen) Is colitis a result of an overreaction of the immune system (defective pro-inflammatory response) to commensal bacteria? (e.g. defects in *NOD2, JAK2*/*STAT3*) or an 'under reaction' of the immune system to a legitimate microbial stimuli? (e.g. a defect in the autophagy pathway; *ATG16L1, IRGM, LRRK*) or, is the altered gut microbiota, a phenomena increasingly observed in patients suffering with IBD, responsible for the onset of the disease? If so, what molecular mechanisms are they implementing in order to do so?

The hypothesis of IBD as almost an autoimmune disease suggests that IBD would occur regardless of the balance of the gut microbiota, as long as it was present. Evidence for this stems from research that has shown that the gut microbiota is largely influenced by

the host's immune phenotype [125, 126] and that colitis has been induced when germ-free animal models of colitis are reconstituted with a 'healthy' microbiota [127]. Additional evidence for this theory come from the effect of that antibiotics administered topically that have been shown to reduce colitis [128]. However, these studies did not utilise next generation sequencing to determine which bacteria were most affected by the antibiotics and for the former argument (colitis developing after introduction of a normal microbiota) it was not determined whether the established gut microbiota was similar to the microbiota that was initially administered and it is widely known that the gut microbiota is integral to development of innate immunity [129]. Consequently, germ-free animals are likely unable to appropriately respond to commensal bacteria regardless of colitis susceptibility genotype and are therefore, not an appropriate model for assessing the impact on the normal gut microbiota on colitis and were not used in this research.

Research into the role of a defective immune response to microbial stimuli, the role of a lone pathogen as mentioned previously and dysbiosis of the gut microbiota in IBD certainly suggest a more direct and specific role for gut microorganisms in the pathogenesis of the disease.

It has been widely observed that dysbiosis of the gut microbiota is associated with IBD as well as many other diseases [63, 130]. 16S rRNA gene analysis and comparison of the two cohorts in this study also revealed significant difference in the gut microbiota of IBD patients compared with healthy volunteers. However it remains to be determined whether this is a cause or an effect of the disease. Although there is mounting evidence to suggest that one of the main causative factors of IBD might be due to a 'shift' in the ratio between probiotic or neutral commensals to more prominent 'aggressive' commensals in IBD [131] which have the capacity to interact with the gut and ultimately; adhere to the mucosal lining, disrupt epithelial barrier integrity, evade the immune response and alter visceral sensitivity and bowel habits [106] The fact that IBD is observed to have a positive correlation with a 'Western' lifestyle which is associated with a general diet, extensive use of antibiotics, a level of hygiene and improved living conditions etc. suggest that this lifestyle may be an initial driving force behind shaping the gut microbiota and consequently, driving an aggressive gut microbiota capable of instigating IBD in those susceptible. Determining the differences between specific functions of the gut microbiota in healthy individuals compared to the functions of the

gut microbiota in IBD may help ascertain how they might be contributing to disease or is integral in determining what the potential long term affects dysbiosis may be having on the guts of individuals suffering with the disease (regardless of whether dysbiosis is cause or effect). It is apt to say that bacteria are integral to the development of IBD but to understand exactly how, the actual molecular mechanisms, occurring in the gut at the epithelial barrier interface rather than just microbial diversity alone need to be studied more definitively.

In this study, a coupling of 16S rRNA gene analysis and a function-driven approach were implemented to begin to unravel the implications of the gut microbiota in IBD. This function-driven approach was focussed on the gut microbial degradome aka the entire proteolytic repertoire of the gut microbiota.

The action of proteolytic activity is only recently becoming recognised as a significant microbial function relating to human disease and therefore much work is still to be done if we are to understand the impact the microbial degradome is having on us as their human hosts [118]. Proteases are utilised by both pathogenic and commensal microorganisms and have even been implicated in the onset of diseases such as IBD. There is mounting evidence to support the theory that microbial proteases are integral to microbial involved in IBD since protease activity appears to be increased in IBD patients compared with healthy samples [132] though most studies have been limited to studying ulcerative colitis alone.

Research by Steck *et al.,* has demonstrated how microbial proteases may contribute to IBD. Their researched showed how *gelE* is capable of degrading the transmembrane glycoprotein E-cadherin, a protein involved in epithelial barrier function. Degradation of this protein results in a loss of barrier function, increased translocation and intestinal inflammation and may be a trigger for onset of IBD [133]. However in practice, since the gut harbours around 10 to 100 trillion bacteria [134] the analysis of only one microorganism may not be so relevant and similarly the degradome of a particular environment should be seen as a system rather than individual proteases since we cannot discount relationships and interactions between them. Therefore in this study, it was attempted to capture the entire faecal degradome as a representation of the distal gut microbiota degradome to pursue the understanding of the role of proteases as they are in

the gut and to compare proteolytic activity between healthy individuals in our cohort (n=11) and individuals with IBD (n=13). The IBD samples had statistically significantly elevated caseinolytic (P = 1.405e-4) and keratinase activity (P = 1.929e-5, figure 7.2.2 A and C) as determined by azo-casein assay and keratin azure assay respectively. Significantly higher caseinolytic activity has been observed in patients with UC and diarrhoeic IBS before [132, 135, 136] suggesting that the microbiota of those suffering with gastrointestinal disorders is enriched with proteolytic bacteria and perhaps a wider substrate specificity since keratin azure was also significantly higher. Keratin azure is a dye-impregnated keratin from sheep's wool often used to determine the keratinolytic capacity of bacteria and fungi. This substrate was chosen as keratin is a major constituent of intestinal epithelial cells and absence or removal of certain keratin types has been associated with the development of colitis and diarrhoea [137] and dye-impregnated feathers, hair or wool have been demonstrated to be useful substrates for assessing keratinase production [138, 139]. These data also suggests that with the high levels of proteolytic bacteria also have an extended functional repertoire i.e. the high level of protease activity may correspond with high levels of different types of proteases present, hence the enhanced ability to degrade keratin.

Collagenase activity is often considered a virulence factor of microorganisms [140]. Although it initially appeared that azo-coll degradation (collagenase activity) was more prevalent in the IBD cohort, the data were not significantly affected by disease state (figure 7.2.2B) and collagenase activity is a feature of a healthy gut microbiota too. This is concurrent with previous research by Pruteanu and colleagues who studied microbial collagenase activity in a culture-dependent study comparing the activity in faecal samples from healthy controls versus patients with IBD and found that activity was found in 25% of samples regardless of disease [141].

It was decided that since the environment of the distal gut has been found to be altered in those suffering with IBD, the effects of different temperature and pH were to be elucidated to determine if the temperature and pH could be having an effect on the bacteria inhabiting the distal gut and therefore, the proteolytic activity. Optimal temperature for FP activity was 37 °C for both cohorts (figure 7.2.3) which was unsurprising since 37 °C is the optimum temperature for growth in the human gut. The pH assay was conducted as previous studies have shown a drop in intestinal luminal pH

in sufferers of IBD [142] and so it could be hypothesised the IBD cohort FP activity would indicate a low pH tolerance. However both cohorts showed significantly higher FP activity at a pH of 8.0 compared to activity at a more acidic or alkaline pH (figure 7.2.3). Host-derived proteases that have been isolated faecal samples such as trypsin and chymotrypsin have been shown to have an optimal pH of around 8.0 similarly; protease positive bacteria isolated from culture-dependent studies commonly have enzymes that operate optimally at pH 8.0. Clearly acid tolerance observed in proteases of the stomach is not replicated in the small and large intestine as exemplified by FP activity observed here.

The use of protease inhibition assays (ProteSEEKER™) revealed which families of proteases were most abundant in each of the samples. This inhibitor set comprises 12 inhibitors that individually target the different families of proteases reversibly or irreversibly. Both cohorts showed significantly reduced activity in the presence of AEBSF, Aprotinin, chymotrypsin inhibitor, EDTA-Na and PMSF and additionally, Antipain dihydrochloride also showed significant inhibition in the IBD cohort but not in the healthy cohort (figure 7.2.5). AEBSF (or 4-(2-Aminoethyl) benzenesulfonyl fluoride hydrochloride) Aprotinin, chymotrypsin and PMSF (or phenylmethanesulfonylfluoride or phenylmethylsulfonyl fluoride) are all serine protease inhibitors with slightly different mechanisms of action.
The significant levels of inhibition of FP activity upon addition of these various serine inhibitors in both cohorts signifies that there are significant quantities of serine proteases in the human distal gut and also may suggest that some of the FP activity observed can be attributed to host serine proteases such as trypsin and chymotrypsin which are secreted in the duodenum from the pancreas. Previous research has investigated the levels of the gastrointestinal trypsin inhibitor; pancreatic secretory trypsin inhibitor (PSTI) in the guts of healthy controls compared with IBD and found significantly reduced levels in affected areas of the colon in patients with both UC and CD [143] and other studies have found significantly higher tryptic faecal protease activity in patients with CD and UC [111, 135, 144]. A particular study by Midtvedt and colleagues observed an inverse correlation between faecal tryptic activity (FTA) and *Bacteroidetes* suggesting that bacteria in this phyla are normally large contributors to controlling the trypsin-like protease pool by further proteolytic breakdown of trypsin [144]. Evidences for microbial breakdown of trypsin also comes from studies in germ-free and

conventional rats whereby germ-free rats have FTA but conventional rats do not [145] and *Bacteroides* species have been shown to degrade trypsin *in vitro* and *in vivo* [144, 146]. Ultimately this demonstrated ability of species of the gut microbiota that are able to degrade trypsin suggests that the healthy gut microbiota has a role in controlling levels of trypsin and related pancreatic proteases in the gut and dysbiosis of the gut microbiota in disorders cause an abundance of bacteria that do not affect trypsin levels and this may have implications in the pathogenesis of IBD. 16S rRNA gene analysis conducted in this research did show a reduction in the *Bacteroidetes* Phylum in the IBD cohort compared to the healthy cohort (figure 7.2.8A) and similarly in the low FP sample group compared to the high FP group (figure 7.2.12) . While this was not a statistically significant, there does appear to be a trend in the correlation between *Bacteroidetes* and low protease activity and perhaps to further explore the relationship between *Bacteroidetes* and protease activity with regards to IBD, a similar study could be conducted encompassing a larger IBD population. Also, serine protease activity was prevalent in the healthy cohort (figure 7.2.5) so serine FP activity cannot be solely attributed to IBD. The theory that elevated trypsin-like serine protease activity is potentially involved in the pathogenesis of IBD also suggests that it is host proteases that are problematic. Host factors have been heavily implicated in the pathogenesis of IBD including trypsin and matrix metalloproteases and studies in other gastrointestinal disorders such as Irritable Bowel Syndrome with diarrhoea have identified trypsin and chymotrypsin-like proteases as a source of elevated FP activity observed in this disorder [111] However, there are bacterial homologs of trypsin [147, 148]. To attempt to differentiate between host and bacterial, as well as other types of proteases (fungi, plant) two approaches were taken. Firstly a ProteaseARREST™ inhibition assay was used with inhibitor cocktails that were specific for bacterial, mammalian, fungal and plant proteases (referred to now as BPI, MPI, FPI and PPI respectively). Secondly, an OptiPrep™ Density gradient column was used to isolate bacterial cells and compare the protease activity after cellular disruption with that of total protein extract and extracellular protein extract.

Bacterial-specific and mammalian specific proteases were the most abundant types of proteases found in both cohorts (figure 7.2.4) which is unsurprising since it is known that bacterial protease activity occurs in faecal samples of both healthy and unhealthy individuals [119] and also host-derived proteases are evident in faecal extracts in healthy

individuals and IBD individuals [118], but bacterial FP activity was significantly more inhibited than mammalian FP activity in the IBD cohort suggesting more bacterial proteases present in the faecal samples of the IBD sufferers. While the plant protease inhibitor cocktail had no effect, the fungal protease inhibitor cocktail significant reduced FP activity in the IBD cohort (figure 7.2.4). When fungal culture was attempted from remaining faecal samples of each individual (which had been stored at -20 °C) using potato dextrose agar and malt agar supplemented with chloramphenicol (40 $\mu g\, ml^{-1}$) and kanamycin (50 $\mu g\, ml^{-1}$) no fungal colonies were observed (data not shown) and no positive bands were observed following an attempt at 18srRNA amplification by PCR (data not shown). This may be indicative of the faecal samples being frozen for too long and the fungi were no longer viable or at levels too low to detect by molecular methods as has been shown before [149].

These data suggest FP activity can be partly attributed to fungal or yeast proteases in the IBD cohort. There is also increasing evidence to implicate fungi IBD [150] and that there is an increase in diversity and composition of fungi in those with IBD [151]. However, there are limitations of this assay as while the contents and concentrations of the inhibitors in each ProteaseARRTEST™ cocktail are not and will not be disclosed by the company, although each inhibitor contains inhibits specific for the organism is question, it could also be assumed that there is overlap between the types of protease inhibitor presents. Most protease inhibitor cocktails deemed specific for bacterial proteases are comprised of; AEBSF, Aprotinin, Bestatin, E-64, Leupeptin, Pepstatin, PMSF and EDTA. Mammalian protease inhibitor and Fungal protease inhibitor cocktails also often include AEBSF, Aprotinin, Bestatin, E-64, Leupeptin, Pepstatin but rarely PMSF; and inhibitor that caused significantly diminished FP activity in both the healthy and IBD cohort (figure 7.2.5). FPI often include 1,10 Phenanthroline monohydrate; a metalloprotease inhibitor of which the mechanism of action is as a chelator of iron, zinc and other divalent metals so perhaps, since the IBD cohort FP activity was significantly affected in the presence of another metalloprotease inhibitor; Phosphoramidon (figure 7.2.5) the loss in activity in the presence of the FPI cocktail maybe as a result of a significant abundance of metalloproteases in the IBD cohort as opposed to an abundance of fungal proteases. PPI cocktails commonly include E-64, pepstain, bestatin and leupeptin; none of which caused significant loss in activity of

either cohort (figure 7.2.5). It is also important to note that the concentrations of each inhibitor are also likely to be different in each cocktail.

An Optiprep™ density gradient was used for isolation bacterial cells and then FP activity was determined before and after cellular disruption by bead beating. It was found that following cellular disruption, significantly higher relative activity (%) was observed (figure 7.2.4B, $P<0.05$) suggesting that intracellular bacterial proteases were contributing to the FP levels observed from previous assays but disrupted bacterial cells exhibited significantly lower FP activity compared to total protein extract and extracellular protein extract ($P<0.05$) and after normalisation of protein, extracellular protein and total protein extract displayed a similar level of activity indicating that secreted proteases are contributing to a significant amount of proteolysis in the gut lumen. What remains to be determined here however, is how much bacterial proteases are contributing to this extracellular proteolysis and how much can be attributed to the host. Also, which other cells in the faecal samples are being disrupted to release protein contributing to total protein FP activity. As previously mentioned Tooth and colleagues studied faecal protease activity in patients with irritable bowel syndrome with diarrhoea and determined that a significant amount of serine FP activity was of host origin [111]. In Tooth's study however, faecal samples were subject to affinity chromatography through a benzamidine-sepharose resin for the specific purification of serine proteases. Thus, while this research of Tooth and colleagues suggests faecal serine proteases are mainly of host origin but this does not infer information reading to the origin of other types of protease which have been found to be present in faeces in this research although the use of affinity chromatography would definitely be a useful direction for future progression of this study and will be discussed in more detail later on.

The results of both protease inhibition assays (figure 7.2.4 and figure 7.2.5) were concurrent with the findings of the general protease activity assays in that a significant proportion of FP activity could be attributed to metallo-, as well as serine proteases (which have previously been discussed) which are the families of proteases that most bacterial keratinases belong to [152], and similarly, the significant difference in FP activity as a result of the addition of the metalloprotease inhibitor; phosphoramidon was observed in the IBD cohort but not in the healthy cohort (figure 7.2.5). To further investigate metalloprotease activity and specific serine protease activity a set of three

fluorogenic substrates were chosen based on the ability to degrade also being associated with virulence (figure 7.2.6) . The substrates were; a synthetic matrix metalloprotease substrate, a neutrophil elastase and a chymotrypsin specific substrate. Matrix metalloproteases have been implicated in IBD [118, 153] and have been assumed to be of host origin. In this study, MMP activity was elevated in the IBD cohort (figure 7.2.5). Thus this data, coupled with the inhibition assay strongly suggest a role for metalloprotease activity in IBD. Studies as part of the IPODD consortium have similarly determined a role for bacterial degradation of extracellular matrix in IBD [141]. Research has also shown metalloproteases previously thought to be of host origin to actually be bacteria-derived [104, 154]. Also, many proteases that are potential virulence factors in the human gut are metalloproteases for example *GelE* from *E. faecalis* and fragilysin from *Bacteroides fragilis.* Serine proteases are also required for pathogenicity in many bacteria [155]. Their presence in the healthy samples indicates that they are not acting as virulence factors in IBD. However neutrophil elastase activity was elevated in the IBD cohort (figure 7.2.4 B) though this is perhaps unsurprising since this enzyme is secreted during inflammation [156]and it is likely that these enzymes are of host origin although very little is known about bacterial neutrophil elastase homologs.

This research has highlighted that commensal FP activity is likely having different implications in the healthy gut compared to the IBD gut. Or the fact that there is elevated metalloprotease activity in the IBD gut is indicative of the presence of metalloprotease producing bacteria in IBD guts which are absent or reduced in the healthy gut. However it still remains to be conclusively determined whether these proteases are of host origin or bacterial. Developing methods to both isolate functionally active proteases and then to determine their origin remains a vital future consideration for this research.

A key molecular feature of IBD is the loss of function of the epithelial barrier and an increase in permeability. To further investigate the role of FP as putative virulence factors in the disease, the ability of the faecal protein extracts from each cohort to alter TER and consequently the functional state of paracellular barrier and also the permeability of a human intestinal cell line were determined. A resultant decrease in TER is indicative of disruption of the tight junctions and deregulated paracellular diffusion [157]. It was found that three of the IBD samples caused a significant decrease

in TER but interestingly, so did one sample from the healthy cohort. Upon the addition of protease inhibitors, no samples showed any reduction in TER indicating that the activity of proteases in these samples were contributing to the drop in TER and that *in situ*, proteases from these samples could be affecting epithelial barrier integrity. Interestingly, the healthy sample that caused a decrease in TER did not cause an increase in permeability of the fluorescein molecular marker. This is likely because the fluorescent marker if a relatively large molecule and TER offers a much more sensitive measurement of barrier integrity as TER measures the ion permeability of cell monolayers and so it is likely that only partial disruption of the junctions may have occurred [157]. IBD samples that caused a significant decrease in TER measurements also caused a significant increase in permeability as did the concentrated supernatant of the *E. faecalis* isolate thus indicating that tight junctions were disrupted to such an extent that large molecules could pass through the cellular monolayer. Surprisingly, one IBD sample that had not caused a decrease in TER appeared to cause an increase in cellular permeability. At this point it can only be speculated that this may be due to the involvement of an intracellular pathway causing endocytosis and exocytosis of the fluorescent marker, but the reasons for this are unclear. These data definitely indicate that FP of some samples are contributing to an altered functional state of the epithelial barrier but as some, but not all samples were capable of this drop in TER and increase in cellular permeability it was hypothesised that this may be again due to elevated levels of protease activity as indeed, the IBD samples that reduced TER fell into the high protease activity bracket of protease activity within the entire cohort. To further validate this, statistical analysis approach was implemented and it was determined that TER was negatively correlated with FP activity i.e. the higher the FP activity, the lower the TER after incubation. This indicates that faecal proteases are a possible contributor to a compromised epithelial barrier and may be a particular problem in IBD.

Illumna Miseq sequencing technology was utilised to compare the bacterial community composition of the distal gut microbiota of patients with IBD with those of the healthy controls and between the two disease pathologies in this study; UC and CD. All samples were dominated by *Firmicutes, Bacteroidetes* and *Proteobacteria* but upon initial analysis of relative proportions of bacterial phyla present in each sample, there were noticeable differences between the healthy and IBD cohort with a marked reduction in

Bacteroidetes and an expansion of *Actinobacteria* and *Proteobacteria* in the IBD cohort (figure 7.2.8A).

Analysis of the lower taxonomic divisions revealed significant differences at the Class, Family and Genus level. While the *Alphaproteobacteria* were significantly increased in the healthy cohort (P = 8.99e-3), the *Gammaproetobacteria* were markedly increased in the IBD cohort (P = 0.025). The *Enterobacteriaceae* family, which belongs to the *Gammaproteobacteria* Class were also significantly increased in the IBD cohort compared to the healthy (P = 0.011 and P = 0.049 respectively). The *Proteobacteria* are increasingly suggested to be implicated in IBD largely due to that fact that this phyla includes the genera *Camplyobacter, Helicobacter, Klebsiella* and the species; *E. coli* which have been associated with IBD [158] and most studies are concurrent with this study in finding an abundance of *Gammaproteobacteria* in IBD groups [159]. However the implications of this have only gone so far as to describe lone pathogens that belong to this class which includes *Klebsiella* and *E. coli* for which their roles in IBD have been discussed previously, but the *Gammaproteobacteria* is a large class that encompasses the second greatest number of genera of all bacteria phyla second only to *Firmicutes* [160]. Members of the *Gammaproteobacteria* are morphologically, biochemically and physiologically diverse also making it difficult to speculate a putative role for their expansion and its association with IBD. Looking more closely at the lower taxonomic divisions that were found to be significantly altered in IBD such as the *Enterobacteriaceae* family may help shed some light upon the role of proteobacteria in IBD. Winter and colleagues have provided evidence that members of the *Enterobacteriaceae* family such as *E. coli* have the genetic capacity to utilise certain by-products of intestinal inflammation such as nitrate, and thus are selected for by the environment created by the inflamed gut. This in turn causes harmful bacteria with this genetic capacity to utilise these products to outcompete those that can't and ultimately this can exacerbate the severity of inflammation [161]. Therefore, perhaps it is fair to suggest that the *Enterobacteriaceae* family have a genetic repertoire that allows them to thrive in a damaged gut and ultimately cause more damage [162].

This study is thus far concurrent with a number of other molecular based studies observing compositional alterations in the gut microbiota in IBD. Frank and colleagues implemented a rRNA sequence analysis approach on gastrointestinal tissue samples to study the diversity of the gut microbiota in IBD and found their data to be dominated by

the four phyla; *Firmicutes*, *Bacteroidetes*, *Proteobacteria* and *Actinobacteria* but with marked reductions in the *Bacteroidetes* in the IBD cohort and substantial increases in *Actinobacteria* and *Proteobacteria* [163]. Similarly, Manichanh and colleagues observed the same dominant phyla as did Mangin and colleagues who also observed a similar distinctive increase in the *Enterobacteria* [164]. It is a noteworthy that these studies also found a distinguishing reduction in *Firmicutes*. When the *Firmicutes* phyla alone were subject to closer scrutiny, there were noticeable differences in the Family taxonomic division between the two cohorts (figure 7.2.8 C). One of the most noticeable differences was in the *Rumminococcaceae* Family ($P = 0.029$) and the *Clostridiaceae* family ($P = 0.048$). Both families include metabolically important species with regards to the gut, two genera have received a lot of attention for their putative protective role against IBD. Many studies have observed a correlation in a decrease in abundance of both the *Faecalibacterium* genera [165] and *Butyricicoccus* [166] in IBD. *Butyricicoccus pullicaecorum* has recently emerged as a potential beneficial commensal significantly reduced in IBD cohorts and has been shown to exert anti-inflammatory effects in rat models of IBD [166]. *Faecalibacterium prausnitzii* is the only species within the *Faecalibacterium* genus and this organism in particular has demonstrated anti-inflammatory properties [167] indicating a significant role in gut homeostasis and anti-inflammatory balance in a healthy state conducted by this organism. The anti-inflammatory effects of such organisms can be partly attributed to their ability to produce short chain fatty acids (SCFAs). In the pinnacle work of Marchesi and colleagues [168] investigating the metabonome in IBD, SCFAs were found to be significantly depleted in IBD which coincided with a depletion of SCFA producing microorganisms. Their work strongly suggests a correlation between levels of SCFAs and dysbiosis of the gut microbiota in IBD as well as providing another non-invasive means of characterising IBD.

While most studies analysing the gut microbiota in IBD have confirmed a reduction in *F. prausnitizii* abundance in IBD [66, 163-165] a recent study by Hansen and colleagues challenged this notion as they confirmed an significant increase in *F. prausnitzii* in paediatric CD patients [169] thus furthering the complexity of an already pathogenically multifarious disease. While animal models of IBD have provided generous amounts of information regarding an anti-inflammatory role for this organism, perhaps the use of *F. prausnitzii* as well as other microorganisms as a putative probiotic in human studies

would be more useful in determining whether they are integral to controlling inflammation or simply innocent bystanders. Particularly, R. Balfour Sartor, an MD and leading researcher in the field of the gut microbiota and IBD has stated an area of research that needs to be developed further is in the coupled use of probiotics and immunosuppressive agents [170].

While this study, as well as copious others have demonstrated a dysbiosis and reduction in diversity within the gut microbiota in IBD, it is important to remember that IBD encompasses numerous disease types, the most commonly occurring being UC and CD. In this study the gut microbiota composition of the UC (n = 6) and CD (n = 7) were also compared. Initial observation of the different dominant Phyla abundance showed some differences, but not as apparent as the entire IBD cohort compared to the healthy cohort (figure 7.2.8 A and B). The UC group were dominated by *Firmicutes* and *Bacteroidetes*, and to a lesser extent; *Actinobacteria* and *Proteobacteria* while the CD group had an abundance of *Proteobacteria* and fewer *Firmicutes* and *Bacteroidetes*, however no significant differences were observed between the two cohorts. Larger scale studies have also found no significant difference in richness and diversity between the two disease types, but also that the pattern of microbial shift is more prevalent in CD than in UC [171]. They have found significant differences in *Proteobacteria* between the two disease types, particularly the *Gammaproteobacteria* [172].
While it is fairly apparent that dysbiosis of the gut microbiota is indeed a factor in IBD, with the increasing number of studies that aim to decipher the implications of dysbiosis of the gut microbiota in IBD , comes an increase in conflicting, as well as supportive data. There are an unprecedented number of variables that affect the pathogenesis of IBD such as diet, lifestyle, host genotype, host immunity and the fact that many participants recruited in studies have different levels of disease severity which all contribute to making the gut microbiota in IBD incredibly difficult to study. This highlights the necessity of implementing functional analysis and ‘omic’ approaches to assess how activity of the gut microbiota might be contributing to IBD.

Since it is hypothesised that shifts in microbial composition in IBD have marked implications on the function of the gut microbiota, particularly with regard to their metabolic capability,

it is fair to perceive that this could be coupled with the increased FP activity in the IBD cohort observed in this study. FP activity was shown to separate depending on high or low FP as revealed by nMDS analysis of the entire cohort, the IBD samples only and the healthy samples (figure 7.2.11) indicating that high FP activity is associated with an alteration in the diversity of the gut microbiota. High FP activity was also associated with a decrease in bacterial diversity (table 7.2.3) a phenomena observed in IBS patients before [173].

To determine any bacterial composition associations with FP activity in IBD, the data were split into quartiles of high and low protease activity and the bacterial communities of each sample group were assessed. Unfortunately, this resulted in a relatively small data set (high FP activity n= 6 and low FP activity n = 6) but it did provide some important preliminary information for future, more large scale studies of gut microbial protease activity and implications in IBD. Initial trends observed were a noticeable reduction in *Proteobacteria* and *Bacteroidetes* in the low FP group (figure 7.2.12) and an increase in *Verrucomicrobia*.

After applying a stringent compositional analysis related to FP activity it was found that the *Verrucomicrobia* were positively associated with low FP activity (P = 0.042). *Verrucomicrobia* is a relatively small phylum with variable molecular and metabolic characteristics but one of the most noteworthy species with regards to the human gut microbiota is *Akkermania muciniphila. A muciniphila* was first isolated as a novel mucin-degrading bacteria from human faeces in the guts of healthy individuals [174] there is also evidence to suggest the species has an intimate relationship with the human mucosal barrier and may exert beneficial functions with regards to gut barrier function and maintenance of gut homeostasis. The organisms can also produce short chain fatty acids as a by-product of mucin degradation [174, 175]. It has also been negatively associated with IBD [176].

Research by both Pullan *et al.,* and Fyderek *et al.,* have observed a significantly diminished mucus layer in patients with IBD compared with healthy controls [177, 178]. While *A. muciniphila* is reduced in IBD, a preponderance of other mucosa-associated bacterial species often considered harmful, have been found to dominate the gut mucosa, including *Streptococcus* spp. [178] and the mucolytic bacterial species; *Ruminococcus gnavus* and *R. torques* [176]. It is yet to be determined whether the loss in species such as *A. muciniphila* and dominance of *R. gnavus* and *R. torques* directly contributes to a diminished mucus layer or whether the converse is true, i.e. a diminished mucus layer

leads to a loss in *A. muciniphila*. *A. muciniphila* has an extended mucinolytic, as well as proteolytic repertoire and indeed this trait appears to be shared by most of the *Verrucomicrobia* [179] as determined by analyses of their genomes [179]. Aside from mucinolytic capacity, the role of these enzymes remains relatively unexplored. Additionally, it is important to note that no difference between mucinase activities between the two cohorts was observed in this study. The decrease in this highly proteolytic phylum in the low FP cohort does lend itself to the hypothesis that this phylum may have a level of control over other proteolytic microorganism and upon its diminution, other highly proteolytic genera are able to dominate, and perhaps this is exerting a deleterious effect on the host. *Verrucomicrobia* are widely considered a beneficial phylum in the gut and have been negatively associated with FP activity here, this, coupled with other evidence from this study that protease activity is elevated in the inflamed gut, it is clear that there is mounting evidence to implicate a role of degradative enzymes of bacteria in the pathogenesis of IBD. Though it remains unclear whether the marked reductions in *Verrucomicrobia* is as a consequence of an inflamed gut or is a possible cause and if it were a cause, what are the driving forces behind loss of *Verrucomicrobia*? Further work needs to be conducted to determine what, if any, the role of species belonging to the *Verrucomicrobia* in inflammation might be, for example it would be very interesting to determine any anti-inflammatory, probiotic or mucosa-thickening effects, of members of the *Verrucomicrobia* as well as their interaction with other bacteria, particularly (opportunistic/putative) pathogens in IBD, *in vitro* and *in vivo* as it is believed currently that this is an avenue yet to be explored.

To conclude, this is the first study to the author's knowledge that has involved a relatively large cohort to conduct a comprehensive analysis coupling function-based characterisation as well as 16S rRNA bacterial community characterisation to explore the faecal degradome in the IBD gut. Distinctions in the microbiota of the healthy cohort compared to the IBD cohort were observed and results were generally concurrent with previous studies conducting 16S rRNA gene profiling of bacterial communities in IBD. Distinctions of the bacterial communities of high FP activity versus low FP activity were additionally observed with high FP activity also characterised by a reduction in microbial diversity. Perhaps the most significant observation was a marked decrease in the *Verrucomicrobia*, the implications of which remain to be explored in future studies.

It has been determined that increased protease activity can be attributed to patients with IBD, in particular, this elevated activity appears to be mainly ascribed to metallo-type proteases. This increase in FP activity may also have consequences for gut physiology and homeostasis in terms of the ability to cause disruption of the gut epithelial barrier as demonstrated here whereby protein extracts were capable of decreasing TER and increasing paracellular permeability. This was an occurrence more characteristic of the IBD cohort. This study therefore indicates that elevated FP activity is likely to be acting as a detrimental factor in IBD and that metalloproteases are the most probable source of the deleterious effects of FP activity as other types of proteases present were also high in the healthy cohort. While the fundamental question of whether protease activity is involved in the progression of IBD or whether it's a consequence of dysbiosis which is in itself, is a consequence, not cause, of the disease, it is certainly clear from results of this study that the gut degradome has the potential to act as a significant virulence factor.

This study has also highlighted that perhaps more efforts should be made to elucidate the origin of FP activity (i.e. host, bacterial, fungal) to determine whether proteases previously thought to be of host origin are actually bacterial which may then offer alternative possibility for therapeutic intervention. Exploration into techniques for isolating proteases from faecal samples in order to functionally characterise them and deduce the sequence would be a priority in progressing this research. One interesting methodology that has potential for protease isolation from faecal samples would be the implementation of affinity chromatography columns incorporated with an immobilised ligand specific for different types of proteases. This methodology has been implemented for isolation of proteases from bacterial cell culture [180] and faecal samples [111]. However, the aims here are to access the entire degradome this research has indicated that the isolation, identification and characterisation of metalloproteases would be more pressing in terms of IBD. Future research could entail the development of a chromatography column with a matrix of metalloprotease specific ligands which will reversibly bind to metalloproteases which could subsequently be eluted and characterised functionally as well as have its amino acid sequence determined in order to hopefully elucidate its origin.

This research has validated that proteolytic activity is a significant factor in the onset of IBD. This lends to the idea that the implementation of protease inhibitors would be a

useful strategy for therapeutic intervention. As described earlier, other research projects have begun to unravel the potential of inhibitors in treating the disease [99-102] with mant studies providing promising results particularly the use of endogenous serine protease inhibitor; elafin [102] . However many of the inhibitors used are very broad acting, such as the use of inhibitors that act as metal ion chelators to halt protease activity [101] and this could have deleterious repercussions for other cellular processes. Also, our research here suggests that elevated protease acitivity may be largely attributed to the gut microbiota and alterations in it's composition lead to increase and decrease in activity. Therefore more research could focus on deciphering protease inhibitors that could be used to target microbial proteases. These protease inhibitors may also be microbial-derived and therefore it would be very useful to begin to unravel the anti-degradome as well as the degradome of the gut microbiota via functional as well as sequence analysis.

Findings of this study provide the basis for much future research particularly in the area of metagenomics. Metagenomics will provide a means of uncovering proteases and their function and to some extent, their origin. However, there is still a lot of novelty in the gut microbiota and many proteases remain uncharacterised. This highlights the importance of the development of novel function-based assay techniques for protease characterisation.

7.4 MATERIALS AND METHODS

7.4.1 Characterisation of faecal proteases within the healthy cohort and the IBD patients cohort

Characterisation of FP activity in each cohort was conducted using a work flow outlined in figure 7.4.1. Faecal samples were collected from volunteers and divided into 1 g samples in sterile tubes and frozen at -20 °C. Protein extraction was conducted as described in Chapter 2.0 and protein concentration was estimated using the BCA assay and normalised to 1 mg/mL for subsequent general protease activity assays using the colorimetric substrates; azo-casein, azo-coll and keratin azure. Procedures were conducted as described in Chapter 2.0. DNA was also extracted from faecal samples following the methods outlined in Chapter 2.0.

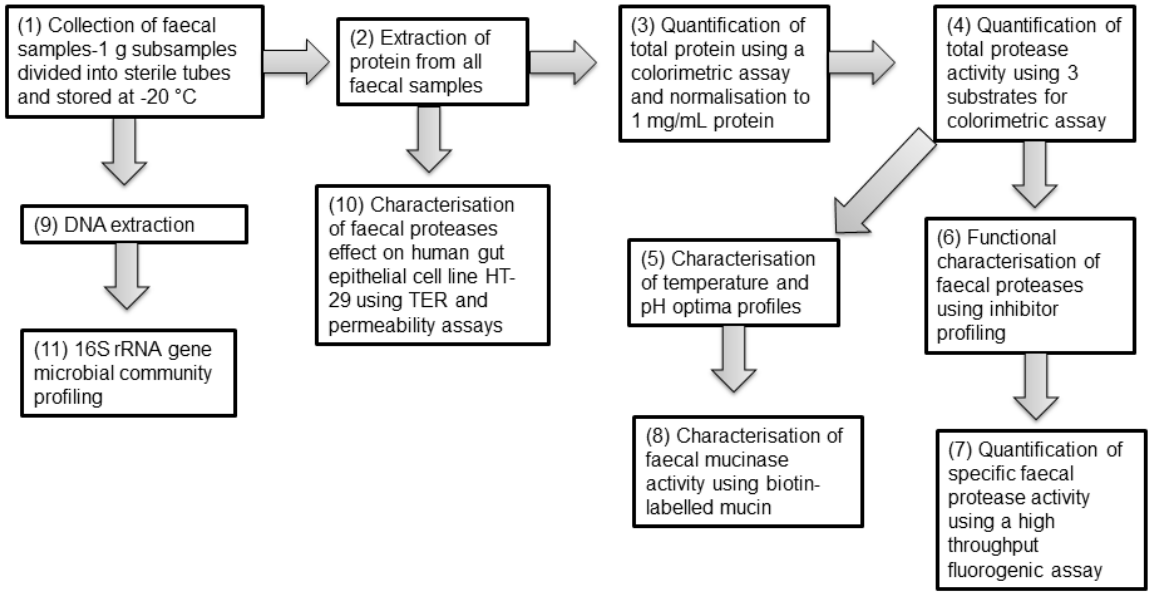


Figure 7.4.1 Brief outline of the methods used and the workflow of this study

7.4.2 Protease inhibitor assays

Protease inhibitor assays to determine the origin of FP activity was conducted using Bacterial, Mammalian, Plant and Yeast ProteaseArrest™ (G-Biosciences, St. Louis, USA) inhibitor cocktails. Assays were carried out according to the manufacturer's instructions on protein extracts from each individual in the study after normalisation to 1 mg/mL protein per sample.
Functional characterisation of the types of proteases present in each sample was conducted using ProteSEEKER™ (G-Biosciences) according to the manufacturer's instructions.

7.4.3 Optiprep™ denisty gradient for isolation of bacterial cells

A 10% w/v faecal suspension was prepared using fresh faeces from 3 samples of the entire cohort in sterile PBS. The suspension was centrifuged at 200 x g for 5 min at 4°C in order to remove faecal debris. The resulting supernatant was collected and subject to centrifugation at 5,500 x g for 20 min at 4°C. The bacterial pellet was resuspended in 4 mL 22% w/v Optiprep™ (Simga-Alrich, Dorset, UK) in Phosphate Buffered Saline (PBS) (pH 7.0). An Optiprep™ density gradient was prepared as follows; 4 mL (0.6 ml) 50% (w/v) OptiPrep™ in PBS was applied to the bottom of the gradient, 4 ml 22% (w/v) crude faecal suspension was very carefully applied over the top and this was over-laid with 4 mL PBS. The gradient was subject to centrifugation at 10,000 x g for 1 hr at 4°C. After such time, the bacterial cell layer, which could be distinctly found above the bottom layer was carefully removed. Any remaining OptiPrep™ solution was removed from the cells by diluting the suspension in PBS and centrifuging at 5,500 x g for 20 min at 4°C. This was carried out a further two times and the cell pellet was resuspended in PBS (pH 7.0). To determine the potential contribution of intracellular proteases to total FP activity, half of each sample was subject to 3 cycles of bead beating to disrupt bacterial cells and release intracellular protein. Protein concentration of each sample was estimated using the BCA assay and subject to protease assay using azo-casein as the substrate.

7.4.4 Protease assays using specific substrates;

Mca-Lys-Pro-Leu-Gly-Leu-Dpa-Ala-Arg-NH_2 was obtained from Enzo Life Sciences (Exeter, UK) and was used as a synthetic fluorogenic substrate for matrix metalloprotease activity. The assay was conducted as previously described [181]. Briefly, MMP substrate was reconstituted in molecular grade water for stock solution (20 mM). This stock solution was further diluted in fluorometric buffer (50 mM Tris, 0.1 M NaCl, , 0.05% (w/v) Brij35, 0.1% (w/v) pH 7.5) to a final concentration of 5 µM. 100 µl 1 mg/mL protein extracts were incubated with the substrate for 2 hours in a 96-well plate. Fluorescence was measured (λ_{ex} = 325 nm, λ_{em} = 395 nm) using the Tecan® Infinite M200 Pro Microplate reader (Tecan®, Zurich, Switzerland). Data was plotted as the mean of triplicate data as relative fluorescence units after subtraction of blank measurements.

MeOSuc-AAPV-AMC was obtained from Enzo Life Sciences (Exeter, UK) as a synthetic, highly sensitive substrate for neutrophil elastase activity. The substrate stock solution was prepared in Dimethyl sulfoxide (DMSO) at a concentration of 10 mM. This was further diluted to 200 µM in fluorometric buffer (50 mM Tris, 0.1 M NaCl, 0.05% (W/v Brij-35, pH 7.5) and incubated for 2 hours in a 96-well plate with 100 µL protein extract from each sample in triplicate. Fluorescence was read (λ_{ex} = 380nm, λ_{em} = 460 nm) using a Tecan® Infinite M200 Pro Microplate reader (Tecan®) and data was plotted as the mean of triplicate data after subtraction from blank measurements.

Suc-Ala-Ala-Pro-Phe-AMC was used as a synthetic fluorogenic substrate for Chymotrypsin-like activity and was obtained from Merck Millipore (Darmstadt, Germany). The assay was conducted as previously described [182]. Briefly, Suc-Ala-Ala-Pro-Phe-AMC was reconstituted in 50 mM Tris-HCl (pH 8.0) to a final concentration of 125 nmol/mL and incubated with 100 µL of 1mg/mL protein extracts for 2 hours at 37 °C. After incubation activity was measured fluorometrically (λ_{ex} = 380 nm, λ_{em} = 460nm) using a Tecan® Infinite M200 Pro Microplate reader (Tecan®).

7.4.5 Biotinylation of mucin

Biotin N-hydroxysuccinimide ester (BNHS) was dissolved in dimethylfomramide (DMF) to a final concentration of 20 mg/mL. 0.1 mL BNHS solution was added to 1 mg

porcine mucin (Sigma-Aldrich) dissolved in 0.9 mL PBS and incubated at 4 °C overnight. The biotin/mucin solution was applied to a Sephadex G25 column and run in PBS. Fractions were collected and subject to slot blot assays to ensure mucin was present. This was carried out by serially diluting fractions (10-fold dilution series) in PBS and adding 5 µL aliquots to a moist nitrocellulose membrane. After drying, the membranes were transferred to freshly prepared periodic acid, acetic acid solution (1% (v/v) and 3% (v/) respectively) and covered for 30 min at room temperature. Membranes were rinsed twice 2 min per wash) ikn freshly prepared 0.1% (w/v) sodium metabisulfate in 1 mM HCl and transferred to Schiff's reagent (Sigma-aldrich) for 15 min. Samples were washed a further 3 times with sodium metabisulfate solution and allowed to dry. Pink-red coloration was indicative of PAS-reactive glycoprotein i.e. mucin.

Fractions containing mucin that had been successfully labelled with biotin were pooled and stored as 0.2 mL samples at 4 °C. Biotin labelled mucin (50 µL) was added to each well of a 96-well plate (Thermo Fisher Scientific, Rockford, IL) and incubated overnight to allow adherence to the plate. Plates were emptied taking care not to disturb the adhered mucin and washed with 10 mM Tris-HCl, pH 8.0 (50 µL) followed by 3 further washing steps (200 µL). Protein extracts were diluted (10-fold) in PBS and triplicate samples of 100 µL of each sample was added to wells of the 96 well plate. The plate was incubated at 37 ˚C for 2 hours. After such time the media was carefully removed and wells were washed 4 times with PBS (200 µL). To block non-specific binding, blocking buffer (1 % Bovine serum albumin in PBS supplemented with 0.2 % (v/v) Tween-20) (200 µL) was added to each well and the microplate was incubated overnight at 4 ˚C. The plates were carefully emptied and washed with PBS (200 µL) twice and incubated with streptavidin-horse radish peroxidase solution (Vector Laboratories, Ltd, Peterborough) (75 µL) for 1 hour at room temperature. The plates were washed 3 times with PBS (200 µL) and incubated in the dark with OPD solution (1,2-phenylenediamine dihydrochloride-Dako UK Ltd, Cambridgeshire, UK) (100 µL) for 1 hour. The reaction was terminated by the addition of 0.5 M sulphuric acid and the resulting absorbance was measured at 490 nm.

7.4.6 MiSeq™ Illumina seqeuncing of 16S rRNA genes

Bacterial community composition of each sample to compare between the two cohorts was characterised by sequencing of the 16S rRNA genes using MiSeq™ Illumina sequencing technology by amplification of the V3 (forward primer = 5'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG) and the V4 (reverse primer =5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC) regions of the 16S rRNA genes by polymerase chain reaction (PCR). These primers were chosen as they had been demonstrated in a previous study to be the most promising primer pair for coverage and phylum spectrum [183].

7.4.7 Analysis of 16S rRNA gene sequences using the Mothur Pipeline

16S rRNA gene data generated by the MiSeq Illumina sequencing platform was processed by the bioinformatics pipeline of mothur [120]. Briefly, reads that were shorter than 200 nucleotides in length, contianed ambiguous bases, chimeric sequences and reads without recognisable primer (forward and reverse) were eliminated. Primers were trimmed from the left over sequences using Mothur. Sequences were aligned to the Silva 16S rRNA sequence reference database [184]. Multiples of the same uniques sequence were removed from to reduce computational power. Sequences were clustered into operational taxonomic units (OTUs) at a level of 97% similarity uisng the SILVA ribosomal RNA gene database [184] and further classified using the Ribosomal database project [185]. Shannon diversity indicies and the chao species richness estimator were conducted using Mothur. STAMP software [121] was used to assess difference between various taxonomic bacterial levels of the two cohorts and also to compare samples with high versus low FP activity using a two-sided Welch's t-test with a Benjamini-Hochberg precedure for mutliple test correction. Heatmaps, Non-metric multi-dimensional scaling (nMDS) anlaysis and significance of diveristy indicies were performed using R software [122] as were all further statistical analyses. The scripts utilised the vegan [186], gplot [187], RcolorBrewer [188] and ggplot 2 [189] packages.

7.4.8 Transepithelial resistance measurments

Pre-warmed culture medium (0.9 mL) was added to each well of a 24-well tissue culture plate (Corning Life Sciences, Lowell, MA). Ascetically, cell culture inserts, 6.5 mm diameter, 0.4 µm pore polyester membrane (Corning Life Sciences) were added to each well and were equilibrated at 37 ºC for 20 min. Cells were seeded at 10^5 cells per sq. cm (0.3 mL volume) onto each insert and cultured under routine conditions (as outlined in Chapter 2X) for 48 hrs by which time a monolayer had developed. After such time, the media from the apical and basolateral compartments was replenished, 100 µL fresh protein extract from each sample (healthy n =11, IBD n = 13) was added to three inserts taking care not to disturb the cell monolayer. The supernatant from a laboratory isolate of *E. faecalis* was added to three inserts as a positive control, PBS was added to a further three inserts and three inserts were left without additional components to act as negative controls. To determine the involvements of proteases on change in TER replica 24-well plates were prepared with the addition of an EDTA-free Halt protease inhibitor cocktail (Thermo Fisher Scientific) supplemented with Phosphoramidon (G-Biosciences).

Trans epithelial resistance measurements (TER) using a voltmeter EVOM^2 (World Precision Instruments, Sarasota, Fl) were taken immediately upon addition of sample and after 24 hrs of incubation. Significance in the change in TER for each sample was determined using the Mann Whitney U test in R software.

7.4.9 Permeability Assay

Cell culture inserts were removed to a new 24-well plate containing fresh culture medium (0.9 mL). To measure translocation, media in the apical compartment was carefully removed and discarded and replaced with Krebs solution (Sigma-Aldrich) containing 500 µg/mL sodium fluorescein (Sigma-Aldrich). The cells were incubated in routine conditions for 2 hrs. Following incubation, 3 X 100 µL aliquots from the outer chamber of each well were removed and transferred to a 96-well plate. Fluorescence was measured (λ_{ex} = 460 nm, λ_{em} = 515nm) using a Tecan® Infinite M200 Pro Microplate reader (Tecan®). Significance was determined using the Mann Whitney U test in R comparing the fluorescence for each sample with and without inhibitor present.

7.5 REFERENCES

1. Hanauer, S.B. (2006). Inflammatory bowel disease: Epidemiology, pathogenesis, and therapeutic opportunities. Inflammatory Bowel Diseases *12*, S3-S9.
2. A framework for human microbiome research. Nature *486*, 215-221.
3. Ghosh, S., and Mitchell, R. (2007). Impact of inflammatory bowel disease on quality of life: Results of the European Federation of Crohn's and Ulcerative Colitis Associations (EFCCA) patient survey. Journal of Crohn's and Colitis *1*, 10-20.
4. Thompson, W.G., Longstreth, G.F., Drossman, D.A., Heaton, K.W., Irvine, E.J., and Műller-Lissner, S.A. (1999). Functional bowel disorders and functional abdominal pain. Gut *45*, II43-II47.
5. Addolorato, G., Capristo, E., Stefanini, G.F., and Gasbarrini, G. (1997). Inflammatory bowel disease: a study of the association between anxiety and depression, physical morbidity, and nutritional status. Scandinavian Journal of Gastroenterology. 1997 Oct;32(10):1013-21.
6. Shanahan, F. (2001). Inflammatory bowel disease: Immunodiagnostics, immunotherapeutics, and ecotherapeutics. Gastroenterology *120*, 622-635.
7. Taxonera, C., Schwartz, D.A., and Garcia-Olmo, D. (2009). Emerging treatments for complex perianal fistula in Crohn's disease. World journal of gastroenterology : WJG *15*, 4263-4272.
8. Danese, S., Sans, M., and Fiocchi, C. (2004). Inflammatory bowel disease: the role of environmental factors. Autoimmunity Reviews *3*, 394-400.
9. Mahid, S.S., Minor, K.S., Soto, R.E., Hornung, C.A., and Galandiuk, S. (2006). Smoking and Inflammatory Bowel Disease: A Meta-analysis. Mayo Clinical Proceedings *81*, 1462-1471.
10. Miller, L.G., Goldstein, G., Murphy, M., and Ginns, L.C. (1982). Reversible alterations in immunoregulatory T cells in smoking. Analysis by monoclonal antibodies and flow cytometry. Chest *82*, 526-529.
11. Cope, G.F., Heatley, R.V., and Kelleher, J.K. (1986). Smoking and colonic mucus in ulcerative colitis. BMJ *293*, 481-481.
12. Koutroubakis, I., Manousos, O.N., Meuwissen, S.G., and Pena, A.S. (1996). Environmental risk factors in inflammatory bowel disease. Hepato-gastroenterology *43*, 381-393.
13. Birrenbach, T., and Bocker, U. (2004). Inflammatory bowel disease and smoking: a review of epidemiology, pathophysiology, and therapeutic implications. Inflammatory Bowel Disease *10*, 848-859.
14. Richardson, C.E., Morgan, J.M., Jasani, B., Green, J.T., Rhodes, J., Williams, G.T., Lindstrom, J., Wonnacott, S., Peel, S., and Thomas, G.A.O. (2003). Effect of smoking and transdermal nicotine on colonic nicotinic acetylcholine receptors in ulcerative colitis. QJM *96*, 57-65.
15. Razani-Boroujerdi, S., Boyd, R.T., Dávila-García, M.I., Nandi, J.S., Mishra, N.C., Singh, S.P., Pena-Philippides, J.C., Langley, R., and Sopori, M.L. (2007). T Cells Express α7-Nicotinic Acetylcholine Receptor Subunits That Require a Functional TCR and Leukocyte-Specific Protein Tyrosine Kinase for Nicotine-Induced Ca2+ Response. The Journal of Immunology *179*, 2889-2898.
16. Sakamoto, N., Kono, S., Wakai, K., Fukuda, Y., Satomi, M., Shimoyama, T., Inaba, Y., Miyake, Y., Sasaki, S., Okamoto, K., et al. (2005). Dietary risk factors

for inflammatory bowel disease: a multicenter case-control study in Japan. Inflammatory Bowel Disease *11*, 154-163.
17. Amre, D.K., D'Souza, S., Morgan, K., Seidman, G., Lambrette, P., Grimard, G., Israel, D., Mack, D., Ghadirian, P., Deslandres, C., et al. (2007). Imbalances in dietary consumption of fatty acids, vegetables, and fruits are associated with risk for Crohn's disease in children. The American journal of gastroenterology *102*, 2016-2025.
18. Persson, P.G., Ahlbom, A., and Hellers, G. (1992). Diet and inflammatory bowel disease: a case-control study. Epidemiology (Cambridge, Mass.) *3*, 47-52.
19. Tragnone, A., Valpiani, D., Miglio, F., Elmi, G., Bazzocchi, G., Pipitone, E., and Lanfranchi, G.A. (1995). Dietary habits as risk factors for inflammatory bowel disease. European journal of gastroenterology & hepatology *7*, 47-51.
20. Russel, M.G., Engels, L.G., Muris, J.W., Limonard, C.B., Volovics, A., Brummer, R.J., and Stockbrugger, R.W. (1998). Modern life' in the epidemiology of inflammatory bowel disease: a case-control study with special emphasis on nutritional factors. Europena Journal of Gastroenterology and Hepatology *10*, 243-249.
21. Hou, J.K., Abraham, B., and El-Serag, H. (2011). Dietary intake and risk of developing inflammatory bowel disease: a systematic review of the literature. The American journal of gastroenterology *106*, 563-573.
22. Mawdsley, J.E., and Rampton, D.S. (2005). Psychological stress in IBD: new insights into pathogenic and therapeutic implications. Gut *54*, 1481-1491.
23. Soderholm, J.D., Yang, P.C., Ceponis, P., Vohra, A., Riddell, R., Sherman, P.M., and Perdue, M.H. (2002). Chronic stress induces mast cell-dependent bacterial adherence and initiates mucosal inflammation in rat intestine. Gastroenterology *123*, 1099-1108.
24. Mazzon, E., Sturniolo, G.C., Puzzolo, D., Frisina, N., and Fries, W. (2002). Effect of stress on the paracellular barrier in the rat ileum. Gut *51*, 507-513.
25. Godet, P.G., May, G.R., and Sutherland, L.R. (1995). Meta-analysis of the role of oral contraceptive agents in inflammatory bowel disease. Gut *37*, 668-673.
26. Cornish, J.A., Tan, E., Simillis, C., Clark, S.K., Teare, J., and Tekkis, P.P. (2008). The risk of oral contraceptives in the etiology of inflammatory bowel disease: a meta-analysis. The American journal of gastroenterology *103*, 2394-2400.
27. Wakefield, A.J., Sawyerr, A.M., Hudson, M., Dhillon, A.P., and Pounder, R.E. (1991). Smoking, the oral contraceptive pill, and Crohn's disease. Digest Diseases Science *36*, 1147-1150.
28. Felder, J.B., Korelitz, B.I., Rajapakse, R., Schwarz, S., Horatagis, A.P., and Gleim, G. (2000). Effects of nonsteroidal antiinflammatory drugs on inflammatory bowel disease: a case-control study. The American journal of gastroenterology *95*, 1949-1954.
29. Cipolla, G., Crema, F., Sacco, S., Moro, E., De Ponti, F., and Frigo, G. (2002). nonsteroidal anti-inflammatory drugs and inflammatory bowel disease: current perspectives. Pharmacological Research *46*, 1-6.
30. Berg, D.J., Zhang, J., Weinstock, J.V., Ismail, H.F., Earle, K.A., Alila, H., Pamukcu, R., Moore, S., and Lynch, R.G. (2002). Rapid development of colitis in NSAID-treated IL-10–deficient mice. Gastroenterology *123*, 1527-1542.
31. MacDermott, R.P. (1994). Alterations in the mucosal immune system in ulcerative colitis and Crohn's disease. The Medical clinics of North America *78*, 1207-1231.

32. Orholm, M., Munkholm, P., Langholz, E., Nielsen, O.H., Sφrensen, T.I.A., and Binder, V. (1991). Familial Occurrence of Inflammatory Bowel Disease. New England Journal of Medicine *324*, 84-88.
33. Brant, S.R. (2011). Update on the heritability of inflammatory bowel disease: The importance of twin studies. Inflammatory Bowel Diseases *17*, 1-5.
34. Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M., et al. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nature Genetics *40*, 955-962.
35. Fisher, S.A., Tremelling, M., Anderson, C.A., Gwilliam, R., Bumpstead, S., Prescott, N.J., Nimmo, E.R., Massey, D., Berzuini, C., Johnson, C., et al. (2008). Genetic determinants of ulcerative colitis include the ECM1 locus and five loci implicated in Crohn's disease. Nature Genetics *40*, 710-712.
36. Rioux, J.D., Xavier, R.J., Taylor, K.D., Silverberg, M.S., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M.M., Datta, L.W., et al. (2007). Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nature Genetics *39*, 596-604.
37. Hugot, J.-P., Chamaillard, M., Zouali, H., Lesage, S., Cezard, J.-P., Belaiche, J., Almer, S., Tysk, C., O'Morain, C.A., Gassull, M., et al. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature *411*, 599-603.
38. Ogura, Y., Bonen, D.K., Inohara, N., Nicolae, D.L., Chen, F.F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerr, R.H., et al. (2001). A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature *411*, 603-606.
39. Rioux, J.D., Silverberg, M.S., Daly, M.J., Steinhart, A.H., McLeod, R.S., Griffiths, A.M., Green, T., Brettin, T.S., Stone, V., Bull, S.B., et al. (2000). Genomewide Search in Canadian Families with Inflammatory Bowel Disease Reveals Two Novel Susceptibility Loci. The American Journal of Human Genetics *66*, 1863-1870.
40. Rioux, J.D., Daly, M.J., Silverberg, M.S., Lindblad, K., Steinhart, H., Cohen, Z., Delmonte, T., Kocher, K., Miller, K., Guschwan, S., et al. (2001). Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nature Genetics *29*, 223-228.
41. Kugathasan, S., Baldassano, R.N., Bradfield, J.P., Sleiman, P.M.A., Imielinski, M., Guthery, S.L., Cucchiara, S., Kim, C.E., Frackelton, E.C., Annaiah, K., et al. (2008). Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. Nature Genetics *40*, 1211-1215.
42. Duerr, R.H., Taylor, K.D., Brant, S.R., Rioux, J.D., Silverberg, M.S., Daly, M.J., Steinhart, A.H., Abraham, C., Regueiro, M., Griffiths, A., et al. (2006). A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene. Science *314*, 1461-1463.
43. Forcione, D.G., Sands, B., Isselbacher, K.J., Rustgi, A., Podolsky, D.K., and Pillai, S. (1996). An increased risk of Crohn's disease in individuals who inherit the HLA class II DRB3*0301 allele. Proceedings of the National Academy of Sciences *93*, 5094-5098.
44. Hampe, J., Franke, A., Rosenstiel, P., Till, A., Teuber, M., Huse, K., Albrecht, M., Mayr, G., De La Vega, F.M., Briggs, J., et al. (2007). A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nature Genetics *39*, 207-211.

45. Parkes, M., Barrett, J.C., Prescott, N.J., Tremelling, M., Anderson, C.A., Fisher, S.A., Roberts, R.G., Nimmo, E.R., Cummings, F.R., Soars, D., et al. (2007). Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nature Genetics *39*, 830-832.
46. Noble, C.L., Nimmo, E.R., Drummond, H., Ho, G.T., Tenesa, A., Smith, L., Anderson, N., Arnott, I.D.R., and Satsangi, J. (2005). The Contribution of OCTN1/2 Variants Within the IBD5 Locus to Disease Susceptibility and Severity in Crohn's Disease. Gastroenterology *129*, 1854-1864.
47. Franke, A., Balschun, T., Sina, C., Ellinghaus, D., Hasler, R., Mayr, G., Albrecht, M., Wittig, M., Buchert, E., Nikolaus, S., et al. (2010). Genome-wide association study for ulcerative colitis identifies risk loci at 7q22 and 22q13 (IL17REL). Nature Genetics *42*, 292-294.
48. Franke, A., McGovern, D.P.B., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nature Genetics *42*, 1118-1125.
49. Anderson, C.A., Boucher, G., Lees, C.W., Franke, A., D'Amato, M., Taylor, K.D., Lee, J.C., Goyette, P., Imielinski, M., Latiano, A., et al. (2011). Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nature Genetics *43*, 246-252.
50. Imielinski, M., Baldassano, R.N., Griffiths, A., Russell, R.K., Annese, V., Dubinsky, M., Kugathasan, S., Bradfield, J.P., Walters, T.D., Sleiman, P., et al. (2009). Common variants at five new loci associated with early-onset inflammatory bowel disease. Nature Genetics *41*, 1335-1340.
51. Khor, B., Gardet, A., and Xavier, R.J. (2011). Genetics and pathogenesis of inflammatory bowel disease. Nature *474*, 307-317.
52. Kumar, A., Takada, Y., Boriek, A., and Aggarwal, B. (2004). Nuclear factor-κB: its role in health and disease. J Mol Med *82*, 434-448.
53. Nielsen, O.H., Kirman, I., Rudiger, N., Hendel, J., and Vainer, B. (2003). Upregulation of interleukin-12 and -17 in active inflammatory bowel disease. Scandinavian Jounral of Gastroenterology *38*, 180-185.
54. Momozawa, Y., Mni, M., Nakamura, K., Coppieters, W., Almer, S., Amininejad, L., Cleynen, I., Colombel, J.F., de Rijk, P., Dewit, O., et al. (2011). Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. Nature Genetics *43*, 43-47.
55. Murthy, A., Li, Y., Peng, I., Reichelt, M., Katakam, A.K., Noubade, R., Roose-Girma, M., DeVoss, J., Diehl, L., Graham, R.R., et al. (2014). A Crohn's disease variant in Atg16l1 enhances its degradation by caspase 3. Nature *506*, 456-462.
56. Glas, J., Seiderer, J., Bues, S., Stallhofer, J., Fries, C., Olszak, T., Tsekeri, E., Wetzke, M., Beigel, F., Steib, C., et al. (2013). IRGM variants and susceptibility to inflammatory bowel disease in the German population. PLoS One *8*, e54338.
57. Liu, Z., and Lenardo, M.J. (2012). The role of LRRK2 in inflammatory bowel disease. Cell Research *22*, 1092-1094.
58. Muise, A.M., Walters, T.D., Glowacka, W.K., Griffiths, A.M., Ngan, B.Y., Lan, H., Xu, W., Silverberg, M.S., and Rotin, D. (2009). Polymorphisms in E-cadherin (CDH1) result in a mis-localised cytoplasmic protein that is associated with Crohn's disease. Gut *58*, 1121-1127.
59. Sabath, E., Negoro, H., Beaudry, S., Paniagua, M., Angelow, S., Shah, J., Grammatikakis, N., Yu, A.S., and Denker, B.M. (2008). Galpha12 regulates

protein interactions within the MDCK cell tight junction and inhibits tight-junction assembly. Journal of Cell Science *121*, 814-824.
60. Scharl, M., Paul, G., Weber, A., Jung, B.C., Docherty, M.J., Hausmann, M., Rogler, G., Barrett, K.E., and McCole, D.F. (2009). Protection of epithelial barrier function by the Crohn's disease associated gene protein tyrosine phosphatase n2. Gastroenterology *137*, 2030-2040.e2035.
61. Fiocchi, C., Battisto, J.R., and Farmer, R.G. (1981). Studies on isolated gut mucosal lymphocytes in inflammatory bowel disease. Detection of activated T cells and enhanced proliferation to *Staphylococcus aureus* and lipopolysaccharides. Digestive Diseases Science *26*, 728-736.
62. Funderburg, N.T., Stubblefield Park, S.R., Sung, H.C., Hardy, G., Clagett, B., Ignatz-Hoover, J., Harding, C.V., Fu, P., Katz, J.A., Lederman, M.M., et al. (2013). Circulating CD4(+) and CD8(+) T cells are activated in inflammatory bowel disease and are associated with plasma markers of inflammation. Immunology *140*, 87-97.
63. Xavier, R.J., and Podolsky, D.K. (2007). Unravelling the pathogenesis of inflammatory bowel disease. Nature *448*, 427-434.
64. Takeda, K., Clausen, B.E., Kaisho, T., Tsujimura, T., Terada, N., Forster, I., and Akira, S. (1999). Enhanced Th1 activity and development of chronic enterocolitis in mice devoid of Stat3 in macrophages and neutrophils. Immunity *10*, 39-49.
65. Fukata, M., Michelsen, K.S., Eri, R., Thomas, L.S., Hu, B., Lukasek, K., Nast, C.C., Lechago, J., Xu, R., Naiki, Y., et al. (2005). Toll-like receptor-4 is required for intestinal response to epithelial injury and limiting bacterial translocation in a murine model of acute colitis. American journal of physiology. Gastrointestinal and liver physiology *288*, G1055-1065.
66. Manichanh, C., Borruel, N., Casellas, F., and Guarner, F. (2012). The gut microbiota in IBD. Nature Reviews Gastroenterology Hepatology *9*, 599-608.
67. Flores, R., Shi, J., Gail, M.H., Ravel, J., and Goedert, J.J.(2012). Assessment of the human faecal microbiota: I. Measurement and reproducibility of selected enzymatic activities. Europena Jounral of Clinical Investigation. 8:848-854.
68. Marchesi, J.R., Holmes, E., Khan, F., Kochhar, S., Scanlan, P., Shanahan, F., Wilson, I.D., and Wang, Y. (2007).Rapid and noninvasive metabonomic characterization of inflammatory bowel disease. Journal of Proteome Research. 2007 Feb;6(2):546-51.
69. Golan, L., Livneh-Kol, A., Gonen, E., Yagel, S., Rosenshine, I., and Shpigel, N.Y. (2009). *Mycobacterium avium paratuberculosis* Invades Human Small-Intestinal Goblet Cells and Elicits Inflammation. Journal of Infectious Diseases *199*, 350-354.
70. Bull, T.J., McMinn, E.J., Sidi-Boumedine, K., Skull, A., Durkin, D., Neild, P., Rhodes, G., Pickup, R., and Hermon-Taylor, J. (2003). Detection and verification of *Mycobacterium avium* subsp. *paratuberculosis* in fresh ileocolonic mucosal biopsy specimens from individuals with and without Crohn's disease. Journal of Clinical Microbiology *41*, 2915-2923.
71. Bernstein, C.N., Nayar, G., Hamel, A., and Blanchard, J.F. (2003). Study of animal-borne infections in the mucosas of patients with inflammatory bowel disease and population-based controls. Journal of Clinical Microbiology *41*, 4986-4990.
72. Barnich, N., Carvalho, F.A., Glasser, A.L., Darcha, C., Jantscheff, P., Allez, M., Peeters, H., Bommelaer, G., Desreumaux, P., Colombel, J.F., et al. (2007).

CEACAM6 acts as a receptor for adherent-invasive *E. coli*, supporting ileal mucosa colonization in Crohn disease. Journal of Clinical Investigation *117*, 1566-1574.

73. Chassaing, B., Rolhion, N., de Vallee, A., Salim, S.Y., Prorok-Hamon, M., Neut, C., Campbell, B.J., Soderholm, J.D., Hugot, J.P., Colombel, J.F., et al. (2011). Crohn disease--associated adherent-invasive *E. coli* bacteria target mouse and human Peyer's patches via long polar fimbriae. Journal of Clinical Investigation *121*, 966-975.
74. Darfeuille-Michaud, A., Boudeau, J., Bulois, P., Neut, C., Glasser, A.L., Barnich, N., Bringer, M.A., Swidsinski, A., Beaugerie, L., and Colombel, J.F. (2004). High prevalence of adherent-invasive *Escherichia coli* associated with ileal mucosa in Crohn's disease. Gastroenterology *127*, 412-421.
75. el-Omar, E., Penman, I., Cruikshank, G., Dover, S., Banerjee, S., Williams, C., and McColl, K.E. (1994). Low prevalence of *Helicobacter pylori* in inflammatory bowel disease: association with sulphasalazine. Gut *35*, 1385-1388.
76. Sonnenberg, A., and Genta, R.M. (2012). Low prevalence of *Helicobacter pylori* infection among patients with inflammatory bowel disease. Alimentary pharmacology & therapeutics *35*, 469-476.
77. Zhang, L., Danon, S.J., Grehan, M., Chan, V., Lee, A., and Mitchell, H. (2005). Natural colonization with Helicobacter species and the development of inflammatory bowel disease in interleukin-10-deficient mice. Helicobacter *10*, 223-230.
78. Hansen, R., Thomson, J.M., Fox, J.G., El-Omar, E.M., and Hold, G.L. (2011). Could Helicobacter organisms cause inflammatory bowel disease? FEMS immunology and medical microbiology *61*, 1-14.
79. Thomson, J.M., Hansen, R., Berry, S.H., Hope, M.E., Murray, G.I., Mukhopadhya, I., McLean, M.H., Shen, Z., Fox, J.G., El-Omar, E., et al. (2011). Enterohepatic helicobacter in ulcerative colitis: potential pathogenic entities? PLoS One *6*, e17184.
80. Man, S.M., Zhang, L., Day, A.S., Leach, S.T., Lemberg, D.A., and Mitchell, H. (2010). *Campylobacter concisus* and other Campylobacter species in children with newly diagnosed Crohn's disease. Inflammatpory Bowel Disease *16*, 1008-1016.
81. Zhang, L., Man, S.M., Day, A.S., Leach, S.T., Lemberg, D.A., Dutt, S., Stormon, M., Otley, A., O'Loughlin, E.V., Magoffin, A., et al. (2009). Detection and Isolation of Campylobacter Species Other than *C. jejuni* from Children with Crohn's Disease. Journal of Clinical Microbiology *47*, 453-455.
82. Man, S.M., Kaakoush, N.O., Leach, S.T., Nahidi, L., Lu, H.K., Norman, J., Day, A.S., Zhang, L., and Mitchell, H.M. (2010). Host attachment, invasion, and stimulation of proinflammatory cytokines by *Campylobacter concisus* and other non-Campylobacter jejuni Campylobacter species. The Journal of infectious diseases *202*, 1855-1865.
83. Hold, G.L., Smith, M., Grange, C., Watt, E.R., El-Omar, E.M., and Mukhopadhya, I. (2014). Role of the gut microbiota in inflammatory bowel disease pathogenesis: what have we learnt in the past 10 years? World journal of gastroenterology : WJG *20*, 1192-1210.
84. Golinska, E., Tomusiak, A., Gosiewski, T., Wiecek, G., Machul, A., Mikolajczyk, D., Bulanda, M., Heczko, P.B., and Strus, M. (2013). Virulence

factors of Enterococcus strains isolated from patients with inflammatory bowel disease. World journal of gastroenterology : WJG *19*, 3562-3572.
85. Nell, S., Suerbaum, S., and Josenhans, C. (2010). The impact of the microbiota on the pathogenesis of IBD: lessons from mouse infection models. Nature Reviews Micro *8*, 564-577.
86. Cleynen, I., Juni, P., Bekkering, G.E., Nuesch, E., Mendes, C.T., Schmied, S., Wyder, S., Kellen, E., Villiger, P.M., Rutgeerts, P., et al. (2011). Genetic evidence supporting the association of protease and protease inhibitor genes with inflammatory bowel disease: a systematic review. PLoS One *6*, e24106.
87. Wu, S., Rhee, K.-J., Zhang, M., Franco, A., and Sears, C.L. (2007). *Bacteroides fragilis* toxin stimulates intestinal epithelial cell shedding and gamma-secretase-dependent E-cadherin cleavage. Journal of Cell Science *120*, 1944-1952.
88. Grys, T.E., Walters, L.L., and Welch, R.A. (2006). Characterization of the *StcE* Protease Activity of *Escherichia coli* O157:H7. Journal of Bacteriology *188*, 4646-4653.
89. Steck, N., Hoffmann, M., Sava, I.G., Kim, S.C., Hahne, H., Tonkonogy, S.L., Mair, K., Krueger, D., Pruteanu, M., Shanahan, F., et al. (2011). Enterococcus faecalis metalloprotease compromises epithelial barrier and contributes to intestinal inflammation. Gastroenterology *141*, 959-971.
90. Hoy, B., Lower, M., Weydig, C., Carra, G., Tegtmeyer, N., Geppert, T., Schroder, P., Sewald, N., Backert, S., Schneider, G., et al. (2010). *Helicobacter pylori HtrA* is a new secreted virulence factor that cleaves E-cadherin to disrupt intercellular adhesion. EMBO reports *11*, 798-804.
91. Brøndsted, L., Andersen, M.T., Parker, M., Jørgensen, K., and Ingmer, H. (2005). The HtrA Protease of *Campylobacter jejuni* Is Required for Heat and Oxygen Tolerance and for Optimal Interaction with Human Epithelial Cells. Applied and Environmental Microbiology *71*, 3205-3212.
92. Henderson, I.R., Czeczulin, J., Eslava, C., Noriega, F., and Nataro, J.P. (1999). Characterization of pic, a secreted protease of *Shigella flexneri* and enteroaggregative *Escherichia coli*. Infection and Immunity *67*, 5587-5596.
93. ChapetonMontes, D., Candela, T., Collignon, A., and Janoir, C. (2011). Localization of the *Clostridium difficile* cysteine protease Cwp84 and insights into its maturation process. Journal of Bacteriology *193*, 5314-5321.
94. Giacaman, R.A., Asrani, A.C., Ross, K.F., and Herzberg, M.C. (2009). Cleavage of protease-activated receptors on an immortalized oral epithelial cell line by *Porphyromonas gingivalis* gingipains. Microbiology *155*, 3238-3246.
95. Nair, S., Milohanic, E., and Berche, P. (2000). ClpC ATPase is required for cell adhesion and invasion of *Listeria monocytogenes*. Infect Immun *68*, 7061-7068.
96. Wu, Z., Milton, D., Nybom, P., Sjo, A., and Magnusson, K.E. (1996). *Vibrio cholerae* hemagglutinin/protease (HA/protease) causes morphological changes in cultured epithelial cells and perturbs their paracellular barrier function. Microbioloy Pathogens *21*, 111-123.
97. Baugh, M.D., Perry, M.J., Hollander, A.P., Davies, D.R., Cross, S.S., Lobo, A.J., Taylor, C.J., and Evans, G.S. (1999). Matrix metalloproteinase levels are elevated in inflammatory bowel disease. Gastroenterology *117*, 814-822.
98. Heuschkel, R.B., MacDonald, T.T., Monteleone, G., Bajaj-Elliott, M., Smith, J.A., and Pender, S.L. (2000). Imbalance of stromelysin-1 and TIMP-1 in the mucosal lesions of children with inflammatory bowel disease. Gut *47*, 57-62.
99. Sykes, A.P., Bhogal, R., Brampton, C., Chander, C., Whelan, C., Parsons, M.E., and Bird, J. (1999). The effect of an inhibitor of matrix metalloproteinases on

colonic inflammation in a trinitrobenzenesulphonic acid rat model of inflammatory bowel disease. Alimentary pharmacology & therapeutics *13*, 1535-1542.
100. Di Sebastiano, P., di Mola, F.F., Artese, L., Rossi, C., Mascetta, G., Pernthaler, H., and Innocenti, P. (2001). Beneficial effects of Batimastat (BB-94), a matrix metalloproteinase inhibitor, in rat experimental colitis. Digestion *63*, 234-239.
101. Medina, C., Videla, S., Radomski, A., Radomski, M., Antolin, M., Guarner, F., Vilaseca, J., Salas, A., and Malagelada, J.R. (2001). Therapeutic effect of phenantroline in two rat models of inflammatory bowel disease. Scandinavian Journal of Gastroenterology *36*, 1314-1319.
102. Motta, J.P., Bermudez-Humaran, L.G., Deraison, C., Martin, L., Rolland, C., Rousset, P., Boue, J., Dietrich, G., Chapman, K., Kharrat, P., et al. (2012). Food-grade bacteria expressing elafin protect against inflammation and restore colon homeostasis. Science Translational Medicine *4*, 158ra144.
103. Forsythe, P., and Kunze, W.A.(2012). Voices from within: gut microbes and the CNS. Cellular and Molecular Life Sciences.
104. Medina, C., Santana, A., Llopis, M., Paz-Cabrera, M.C., Antolin, M., Mourelle, M., Guarner, F., Vilaseca, J., Gonzalez, C., Salas, A., et al. (2005). Induction of colonic transmural inflammation by *Bacteroides fragilis*: implication of matrix metalloproteinases. Inflamm Bowel Dis *11*, 99-105.
105. Carroll, I.M., and Maharshak, N. (2013). Enteric bacterial proteases in inflammatory bowel disease- pathophysiology and clinical implications. World Journal of Gastroenterology. *19*, 7531-7543.
106. Steck, N., Mueller, K., Schemann, M., and Haller, D. (2013). Republished: Bacterial proteases in IBD and IBS. Postgraduate Medical Journal *89*, 25-33.
107. Kida, Y., Higashimoto, Y., Inoue, H., Shimizu, T., and Kuwano, K. (2008). A novel secreted protease from *Pseudomonas aeruginosa* activates NF-κB through protease-activated receptors. Cellular Microbiology *10*, 1491-1504.
108. Bueno, L., and Fioramonti, J. (2008). Protease-activated receptor 2 and gut permeability: a review. Neurogastroenterology & Motility *20*, 580-587.
109. Kim, J.A., Choi, S.C., Yun, K.J., Kim, D.K., Han, M.K., Seo, G.S., Yeom, J.J., Kim, T.H., Nah, Y.H., and Lee, Y.M. (2003). Expression of protease-activated receptor 2 in ulcerative colitis. Inflammatory Bowel Disease *9*, 224-229.
110. Kida, Y., Inoue, H., Shimizu, T., and Kuwano, K. (2007). *Serratia marcescens* serralysin induces inflammatory responses through protease-activated receptor 2. Infection and Immunity *75*, 164-174.
111. Tooth, D., Garsed, K., Singh, G., Marciani, L., Lam, C., Fordham, I., Fields, A., Banwait, R., Lingaya, M., Layfield, R., et al. (2014). Characterisation of faecal protease activity in irritable bowel syndrome with diarrhoea: origin and effect of gut transit. Gut *63*, 753-760.
112. Heck, L.W., Alarcon, P.G., Kulhavy, R.M., Morihara, K., Russell, M.W., and Mestecky, J.F. (1990). Degradation of IgA proteins by *Pseudomonas aeruginosa* elastase. Journal of immunology (Baltimore, Md. : 1950) *144*, 2253-2257.
113. Walker, K.E., Moghaddame-Jafari, S., Lockatell, C.V., Johnson, D., and Belas, R. (1999). ZapA, the IgA-degrading metalloprotease of *Proteus mirabilis*, is a virulence factor expressed specifically in swarmer cells. Molecular Microbiology *32*, 825-836.
114. Mikolajczyk-Pawlinska, J., Travis, J., and Potempa, J. (1998). Modulation of interleukin-8 activity by gingipains from *Porphyromonas gingivalis* Implications for pathogenicity of periodontal disease. FEBS Letters *440*, 282-286.

115. Kapur, V., Majesky, M.W., Li, L.L., Black, R.A., and Musser, J.M. (1993). Cleavage of interleukin 1 beta (IL-1 beta) precursor to produce active IL-1 beta by a conserved extracellular cysteine protease from Streptococcus pyogenes. Proceedings of the National Academy of Sciences *90*, 7676-7680.
116. Schultsz, C., van den Berg, F.M., ten Kate, F.W., Tytgat, G.N.J., and Dankert, J. (1999). The intestinal mucus layer from patients with inflammatory bowel disease harbors high numbers of bacteria compared with controls. Gastroenterology *117*, 1089-1097.
117. Hughes, R., Magee, E.A., and Bingham, S. (2000). Protein degradation in the large intestine: relevance to colorectal cancer. Current Issues in Intestinal Microbiology *1*, 51-58.
118. Steck, N., Mueller, K., Schemann, M., and Haller, D. (2013). Bacterial proteases in IBD and IBS. Gut.
119. Macfarlane, G.T., Cummings, J.H., and Allison, C. (1986). Protein Degradation by Human Intestinal Bacteria. Journal of General Microbiology *132*, 1647-1656.
120. Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., et al. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Applied and Environmental Microbiology *75*, 7537-7541.
121. Parks, D.H., and Beiko, R.G. (2010). Identifying biologically relevant differences between metagenomic communities. Bioinformatics *26*, 715-721.
122. Team, R.C. (2012). R: A language and environment for statistical computing.
123. Mizoguchi, A., and Mizoguchi, E. (2010). Animal models of IBD: linkage to human disease. Current opinion in pharmacology *10*, 578-587.
124. Sellon, R.K., Tonkonogy, S., Schultz, M., Dieleman, L.A., Grenther, W., Balish, E., Rennick, D.M., and Sartor, R.B. (1998). Resident enteric bacteria are necessary for development of spontaneous colitis and immune system activation in interleukin-10-deficient mice. Infection and Immunity *66*, 5224-5231.
125. Tamboli, C.P., Neut, C., Desreumaux, P., and Colombel, J.F. (2004). Dysbiosis in inflammatory bowel disease. Gut *53*, 1-4.
126. Dombrowicz, D., Nutten, S., Desreumaux, P., Neut, C., Torpier, G., Peeters, M., Colombel, J.F., and Capron, M. (2001). Role of the high affinity immunoglobulin E receptor in bacterial translocation and intestinal inflammation. The Journal of experimental medicine *193*, 25-34.
127. Kühn, R., Löhler, J., Rennick, D., Rajewsky, K., and Müller, W. (1993). Interleukin-10-deficient mice develop chronic enterocolitis. Cell *75*, 263-274.
128. Guarner, F., and Malagelada, J.-R. (2003). Role of bacteria in experimental colitis. Best Practice & Research Clinical Gastroenterology *17*, 793-804.
129. Hooper, L.V. (2004). Bacterial contributions to mammalian gut development. Trends in Microbiology *12*, 129-134.
130. Spor, A., Koren, O., and Ley, R. (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. Nature Reviews Microbiology *9*, 279-290.
131. Sartor, R.B. (2001). Intestinal microflora in human and experimental inflammatory bowel disease. Current opinion in gastroenterology *17*, 324-330.
132. Annaházi, A., Gecse, K., Dabek, M., Ait-Belgnaoui, A., Rosztóczy, A.s., Ròka, R.r., Molnàr, T.s., Theodorou, V., Wittmann, T., Bueno, L., et al. (2009). Fecal proteases from diarrheic-IBS and ulcerative colitis patients exert opposite effect on visceral sensitivity in mice. PAIN *144*, 209-217.

133. Steck, N., Hoffmann, M., Sava, I.G., Kim, S.C., Hahne, H., Tonkonogy, S.L., Mair, K., Krueger, D., Pruteanu, M., Shanahan, F., et al. (2011).*Enterococcus faecalis* Metalloprotease Compromises Epithelial Barrier and Contributes to Intestinal Inflammation. Gastroenterology *141*, 959-971.
134. Gill, S.R., Pop, M., DeBoy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., Gordon, J.I., Relman, D.A., Fraser-Liggett, C.M., and Nelson, K.E. (2006). Metagenomic analysis of the human distal gut microbiome. Science *312*, 1355-1359.
135. Bustos, D., Negri, G., De Paula, J.A., Di Carlo, M., Yapur, V., Facente, A., and De Paula, A. (1998). Colonic proteinases: increased activity in patients with ulcerative colitis. Medicina *58*, 262-264.
136. Gecse, K., Róka, R., Ferrier, L., Leveque, M., Eutamene, H., Cartier, C., Ait-Belgnaoui, A., Rosztóczy, A., Izbéki, F., Fioramonti, J., et al. (2008). Increased faecal serine protease activity in diarrhoeic IBS patients: a colonic lumenal factor impairing colonic permeability and sensitivity. Gut *57*, 591-599.
137. Toivola, D.M., Krishnan, S., Binder, H.J., Singh, S.K., and Omary, M.B. (2004). Keratins modulate colonocyte electrolyte transport via protein mistargeting. J Cellular Biology *164*, 911-921.
138. Cai, C.-g., Lou, B.-g., and Zheng, X.-d. (2008). Keratinase production and keratin degradation by a mutant strain of *Bacillus subtilis*. Journal of Zhejiang University Science. B *9*, 60-67.
139. Gupta, R., and Ramnani, P. (2006). Microbial keratinases and their prospective applications: an overview. Applied Microbiology and Biotechnology *70*, 21-33.
140. Harrington, D.J. (1996). Bacterial collagenases and collagen-degrading enzymes and their potential role in human disease. Infection and Immunity *64*, 1885-1891.
141. Pruteanu, M., Hyland, N.P., Clarke, D.J., Kiely, B., and Shanahan, F. (2011). Degradation of the extracellular matrix components by bacterial-derived metalloproteases: implications for inflammatory bowel diseases. Inflammatory Bowel Disease *17*, 1189-1200.
142. Nugent, s.g., kumar, d., rampton, d.s., and evans, d.f. (2001). Intestinal luminal pH in inflammatory bowel disease: possible determinants and implications for therapy with aminosalicylates and other drugs. Gut *48*, 571-577.
143. Playford, R.J., Hanby, A.M., Patel, K., and Calam, J. (1995). Influence of inflammatory bowel disease on the distribution and concentration of pancreatic secretory trypsin inhibitor within the colon. The American journal of pathology *146*, 310-316.
144. Midtvedt, T., Zabarovsky, E., Norin, E., Bark, J., Gizatullin, R., Kashuba, V., Ljungqvist, O., Zabarovska, V., Mollby, R., and Ernberg, I. (2013). Increase of faecal tryptic activity relates to changes in the intestinal microbiome: analysis of Crohn's disease with a multidisciplinary platform. PLoS One *8*, e66074.
145. Norin, K.E., Gustafsson, B.E., and Midtvedt, T. (1986). Strain differences in faecal tryptic activity of germ-free and conventional rats. Laboratory Animals *20*, 67-69.
146. Ramare, F., Hautefort, I., Verhe, F., Raibaud, P., and Iovanna, J. (1996). Inactivation of tryptic activity by a human-derived strain of Bacteroides distasonis in the large intestines of gnotobiotic rats and mice. Applied Environmental Microbiology *62*, 1434-1436.
147. Lee, W.S., Park, C.H., and Byun, S.M. (2004). *Streptomyces griseus* Trypsin Is Stabilized against Autolysis by the Cooperation of a Salt Bridge and Cation-π Interaction. Journal of Biochemistry *135*, 93-99.

148. Grenier, D. (1995). Characterization of the trypsin-like activity of *Bacteroides forsythus*. Microbiology *141*, 921-926.
149. Scanlan, P.D., and Marchesi, J.R. (2008). Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and -independent analysis of faeces. ISME J *2*, 1183-1193.
150. Underhill, D., and Braun, J. (2008). Current understanding of fungal microflora in inflammatory bowel disease pathogenesis. Inflammatory Bowel Disease *14*, 1147-1153.
151. Ott, S.J., Kuhbacher, T., Musfeldt, M., Rosenstiel, P., Hellmig, S., Rehman, A., Drews, O., Weichert, W., Timmis, K.N., and Schreiber, S. (2008). Fungi and inflammatory bowel diseases: Alterations of composition and diversity. In Scandinavian Journal of Gastroenterology, Volume 43. (Norway), pp. 831-841.
152. Brandelli, A. (2008). Bacterial Keratinases: Useful Enzymes for Bioprocessing Agroindustrial Wastes and Beyond. Food Bioprocess Technol *1*, 105-116.
153. Ravi, A., Garg, P., and Sitaraman, S.V. (2007). Matrix metalloproteinases in inflammatory bowel disease: boon or a bane? Inflammatory Bowel Disease *13*, 97-107.
154. Medina, C., and Radomski, M.W. (2006). Role of Matrix Metalloproteinases in Intestinal Inflammation. Journal of Pharmacology and Experimental Therapeutics *318*, 933-938.
155. Tripathi, L.P., and Sowdhamini, R. (2008). Genome-wide survey of prokaryotic serine proteases: analysis of distribution and domain architectures of five serine protease families in prokaryotes. BMC Genomics *9*, 549-549.
156. Belaaouaj, A., Kim, K.S., and Shapiro, S.D. (2000). Degradation of outer membrane protein A in Escherichia coli killing by neutrophil elastase. Science *289*, 1185-1188.
157. Matter, K., and Balda, M.S. (2003). Functional analysis of tight junctions. Methods *30*, 228-234.
158. Mukhopadhya, I., Hansen, R., El-Omar, E.M., and Hold, G.L. (2012). IBD[mdash]what role do Proteobacteria play? Nature Reviews Gastroenterology and Hepatology *9*, 219-230.
159. Sartor, R.B., and Mazmanian, S.K. (2012). Intestinal Microbes in Inflammatory Bowel Diseases. Am J Gastroenterol Suppl *1*, 15-21.
160. Williams, K.P., Gillespie, J.J., Sobral, B.W., Nordberg, E.K., Snyder, E.E., Shallom, J.M., and Dickerman, A.W. (2010). Phylogeny of gammaproteobacteria. Journal of Bacteriology *192*, 2305-2314.
161. Winter, S.E., Winter, M.G., Xavier, M.N., Thiennimitr, P., Poon, V., Keestra, A.M., Laughlin, R.C., Gomez, G., Wu, J., Lawhon, S.D., et al. (2013). Host-derived nitrate boosts growth of E. coli in the inflamed gut. Science *339*, 708-711.
162. Winter, S.E., Lopez, C.A., and Baumler, A.J. (2013). The dynamics of gut-associated microbial communities during inflammation. EMBO reports *14*, 319-327.
163. Frank, D.N., St. Amand, A.L., Feldman, R.A., Boedeker, E.C., Harpaz, N., and Pace, N.R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proceedings of the National Academy of Sciences *104*, 13780-13785.
164. Mangin, I., Bonnet, R., Seksik, P., Rigottier-Gois, L., Sutren, M., Bouhnik, Y., Neut, C., Collins, M.D., Colombel, J.-F., Marteau, P., et al. (2004). Molecular

inventory of faecal microflora in patients with Crohn's disease. FEMS Microbiology Ecology *50*, 25-36.
165. Joossens, M., Huys, G., Cnockaert, M., De Preter, V., Verbeke, K., Rutgeerts, P., Vandamme, P., and Vermeire, S. (2011). Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. Gut *60*, 631-637.
166. Eeckhaut, V., Machiels, K., Perrier, C., Romero, C., Maes, S., Flahou, B., Steppe, M., Haesebrouck, F., Sas, B., Ducatelle, R., et al. (2013). *Butyricicoccus pullicaecorum* in inflammatory bowel disease. Gut *62*, 1745-1752.
167. Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermudez-Humaran, L.G., Gratadoux, J.-J., Blugeon, S., Bridonneau, C., Furet, J.-P., Corthier, G., et al. (2008). *Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. Proceedings of the National Academy of Science U S A *105*, 16731-16736.
168. Marchesi, J.R., Holmes, E., Khan, F., Kochhar, S., Scanlan, P., Shanahan, F., Wilson, I.D., and Wang, Y. (2007). Rapid and noninvasive metabonomic characterization of inflammatory bowel disease. Journal of proteome research *6*, 546-551.
169. Hansen, R., Russell, R.K., Reiff, C., Louis, P., McIntosh, F., Berry, S.H., Mukhopadhya, I., Bisset, W.M., Barclay, A.R., Bishop, J., et al. (2012). Microbiota of De-Novo Pediatric IBD: Increased *Faecalibacterium Prausnitzii* and Reduced Bacterial Diversity in Crohn's But Not in Ulcerative Colitis. The American journal of gastroenterology *107*, 1913-1922.
170. Sartor, R.B. (2011). Efficacy of probiotics for the management of inflammatory bowel disease. Gastroenterology & hepatology *7*, 606-608.
171. Wills, E.S., Jonkers, D.M., Savelkoul, P.H., Masclee, A.A., Pierik, M.J., and Penders, J. (2014). Fecal microbial composition of ulcerative colitis and Crohn's disease patients in remission and subsequent exacerbation. PLoS One *9*, e90981.
172. Gophna, U., Sommerfeld, K., Gophna, S., Doolittle, W.F., and Veldhuyzen van Zanten, S.J.O. (2006). Differences between Tissue-Associated Intestinal Microfloras of Patients with Crohn's Disease and Ulcerative Colitis. Journal of Clinical Microbiology *44*, 4136-4141.
173. Carroll, I.M., Ringel-Kulka, T., Ferrier, L., Wu, M.C., Siddle, J.P., Bueno, L., and Ringel, Y. (2013). Fecal protease activity is associated with compositional alterations in the intestinal microbiota. PLoS One *8*, e78017.
174. Derrien, M., Vaughan, E.E., Plugge, C.M., and de Vos, W.M. (2004). *Akkermansia muciniphila* gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. International Journal of Systematic and Evolutionary Microbiology *54*, 1469-1476.
175. Everard, A., Belzer, C., Geurts, L., Ouwerkerk, J.P., Druart, C., Bindels, L.B., Guiot, Y., Derrien, M., Muccioli, G.G., Delzenne, N.M., et al. (2013). Cross-talk between *Akkermansia muciniphila* and intestinal epithelium controls diet-induced obesity. Proceedings of the National Academy of Sciences *110*, 9066-9071.
176. Png, C.W., Linden, S.K., Gilshenan, K.S., Zoetendal, E.G., McSweeney, C.S., Sly, L.I., McGuckin, M.A., and Florin, T.H. (2010). Mucolytic bacteria with increased prevalence in IBD mucosa augment in vitro utilization of mucin by other bacteria. The American journal of gastroenterology *105*, 2420-2428.
177. Pullan, R.D., Thomas, G.A., Rhodes, M., Newcombe, R.G., Williams, G.T., Allen, A., and Rhodes, J. (1994). Thickness of adherent mucus gel on colonic mucosa in humans and its relevance to colitis. Gut *35*, 353-359.

178. Fyderek, K., Strus, M., Kowalska-Duplaga, K., Gosiewski, T., Wedrychowicz, A., Jedynak-Wasowicz, U., Sladek, M., Pieczarkowski, S., Adamski, P., Kochan, P., et al. (2009). Mucosal bacterial microflora and mucus layer thickness in adolescents with inflammatory bowel disease. World journal of gastroenterology : WJG *15*, 5287-5294.
179. van Passel, M.W., Kant, R., Zoetendal, E.G., Plugge, C.M., Derrien, M., Malfatti, S.A., Chain, P.S., Woyke, T., Palva, A., de Vos, W.M., et al. (2011). The genome of *Akkermansia muciniphila*, a dedicated intestinal mucin degrader, and its use in exploring intestinal metagenomes. PLoS One *6*, e16876.
180. Eijsink, V.G., van den Burg, B., and Venema, G. (1991). High performance affinity chromatography of Bacillus neutral proteases. Biotechnology and applied biochemistry *14*, 275-283.
181. Neumann, U., Kubota, H., Frei, K., Ganu, V., and Leppert, D. (2004). Characterization of Mca-Lys-Pro-Leu-Gly-Leu-Dpa-Ala-Arg-NH2, a fluorogenic substrate with increased specificity constants for collagenases and tumor necrosis factor converting enzyme. Analytical Biochemistry *328*, 166-173.
182. Oshima, G. (1983). Stimulation by Phospholipid Vesicles of Proteolysis of Egg White Lysozyme by Chymotrypsin. Journal of Biochemistry *94*, 1615-1620.
183. Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., and Glockner, F.O. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Research *41*, e1.
184. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research *41*, D590-D596.
185. Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., and Tiedje, J.M. (2013). Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Research.
186. Oksanen, J. (2011). Multivariate analysis of ecological communities in R: vegan tutorial. R package version *1*.
187. Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., and Moeller, S. (2013). gplots: Various R programming tools for plotting data. R package version 2.11. 0. (Available online at: h ttp://CRAN. R-project. org/package= gplots).
188. Neuwirth, E. (2007). RColorBrewer: ColorBrewer palettes. R package version 1.0-2.
189. Wilkinson, L. (2011). ggplot2: Elegant Graphics for Data Analysis by Wickham, H. Biometrics *67*, 678-679.

8.0 GENERAL CONCLUSIONS AND DISCUSSION

Comprehensive studies to investigate tools for functional metagenomic isolation of proteases as well as examining the role of microbial proteases in a diseased gut were undertaken. This research has both ameliorated and highlighted further some of the difficulties of FM particularly when screening for proteases. It has also validated that proteases are associated with diseases of the gut, specifically, IBD.

In this chapter, the general conclusions of each main area of work will be reiterated with respect to the main aims and objectives that were explained in Chapter 1.0. In addition, the implications of this study will be discussed in a broader context particularly regarding proteases within the gut microbiota and functional metagenomics along with the future directions that each area of work should be striving towards.

8.1 CONCLUSIONS

1) Aim: To develop a simple, yet robust agar to eliminate the occurrence of false positive in functional metagenomic library screening for proteases (Chapter 3.0).

It was hypotheisised that the use of SMA to screen for protease activity may lead to false positives due to acid production as a result of galactosidase enzyme activity and as a consequence, an alternative simple agar screen was needed. The use of a lactose free skimmed milk agar was explored as this could potentially eliminate the problem of false positive due to galactosidase activity. Indeed, strains and FM isolates previously deemed positive for protease secretion and activity using SMA, but negative using other means were also negative using LFSMA ,but gave a positive protease phenotype when screening known protease producers such as *B. subtilis* MY2016. This Valio™ lactose-free milk agar is therefore an effective and robust agar for correctly identifying proteases via FM as well as for assessing protease acitivity in pure cultures.

2) Aim: To identify and functionally characterise the active protease gene(s) from a fosmid vector isolated using metagenomics from an diary waste treatment plant (Chapter 4.0).

It had initially been hypothesised that activated sludge from a dairy waste treatment would be a bacterial ecosystem enriched for bacteria that have the ability to degrade the protein in diary waste stream and would therefore provide a good sampling site for DNA extraction for FM library construction and screening for bacterial protease activity. This hypothesis proved correct as one large insert (38 kb) fosmid clone was found to confer a proteolytic phenotype. Further analysis of the sequenced insert revealed two putative protease sequences both of which appeared to be able to be responsible for the protease phenotype observed by the fosmid clone. Upon sub-cloning of each putative sequence, however, only one sequence (designated M1-2) was able to recapitulate the proteolytic phenotype even though sequence analysis revealed little reason as to why the other sequence (designated M1-1) did not express a protease. Thus, this suggested that either this sequence did not encode a protease gene, highlighting the negative aspects of *in silico* data mining for bacterial enzymes based on sequence homologies. The lack of expression also suggested that perhaps *E. coli* was just unable to express the gene and

led us to believe that alternative hosts might be more appropriate for creating an FM library for proteases. This research also highlighted that our study also revealed that the genetic locus isolated that encoded M1-2 was also conserved with a well-characterised family of proteases and shared 56 % sequence identity to another metagenome-derived metalloprotease. Further analysis of other studies that had isolated proteases from diverse habitats using FM and claimed novelty uncovered that they too were closely related to already characterised families of proteases as defined by the MEROPS database [1] and the protein sequences shared greater than 50 % homology with their closest matches after BLASTp searches. Really, a new or novel gene and function needs to have no homologue in the database or it should define a cluster of genes annotated as unknown. Thus this prompted the query as to whether or not these proteases could really be called novel and whether or not FM was capable of isolating novel proteases.

3) Aim: To conduct a culture-dependent analysis on protease producers within the gut microbiota (Chapter 5.0).

In this research, one of the overriding objectives was to begin to uncover the role of bacterial proteases in the human gut. One of the first steps towards achieving this goal was to culture and identify gut bacteria that secreted proteases as although the cultivable microorganisms may not be numerically dominant, this may not necessarily imply that their proteases are insignificant. Aerobic protease producers was found to be dominated by the *Enterococcus* genus, particularly *E. faecalis* as well as a *Lactobacillus casei* a species widely considered to be a probiotic. Anaerobic protease producers were likely to belong to the *Bacteroides* and *Bifidobacteria* genera. To the authors knowledge, only one other study has conducted a culture-dependent study on protease producers in the human gut [2] in similar findings were reported except *Clostridia*, *Staphylococci* and *Propionobacteria* were also isolated although they were identified based on morphology and Gram testing. 16S rRNA gene sequencing is arguably a more accurate means of identifying bacteria and so perhaps it is fair to conclude that the study here is more useful and it is likely that *Enterococcus*, *Lactobacillus*, *Bacteroides* and *Bifidobacteria* are the most dominant genera of bacteria that produce proteases in the gut, that we are currently able to culture.

4) Aim: To develop novel metagenomic tools to obtain characterised proteases from (gut) bacteria and to implement these tools to screen the gut microbiota for protease activity (Chapter 5.0).

It was hypothesised that the use of a novel shuttle vector system utilising a Gram negative as well as a Gram positive host that has a demonstrated capacity to express and secrete proteases effectively will be much more efficient in capturing protease genes from the gut metagenome. The Gram positive host; *B. subtilis* WB800N was capable of expressing and secreting two well-characterised proteases from other bacteria while the *E. coli* host could not. This data showed that when attempting a FM screen, host selection is very important and careful thought should be put into choosing a host that has the genetic machinery capable of expressing the target gene. It was also found that one protease gene could not be expressed by ether host and further literature analysis revealed this is because expression of the gene is reliant on a number of other genes too suggesting larger inserts are needed for the expression of some proteases.

Unfortunately, when this host was implemented to screen the gut microbiota for proteases, the transformation procedure was time consuming, and successful transformation was unpredictable. The inadequate success of transformation of *B. subtilis* WB800N resulted in a limited library with regards to genetic coverage due to too few clones compared to the number of clones needed in most successful FM libraries. This research therefore, has highlighted the difficulties associated with screening for proteases using FM strategies. However, from these results, we are able to develop new hypotheses and techniques that will aid future projects and help us to add novelty to the current protease sequence space. Exploration of alternative hosts is one option, particularly with a well-characterised procedure for transformation to allow the generation of enough libraries for adequate coverage of the mgDNA from the environment in question. However, it seems more likely the main issues experienced here are due to the library preparation itself and more efforts should be implemented to ensure fragments of DNA that are >3 kb are used to generate the library to hopefully capture the protease gene as well as any regulatory genes that are also needed for expression [3]. Methodologies to enable this have been discussed in chapter 5.0. Additionally, the methods used in chapter 5.0 did not include recombinant plasmid selection which may have further reduced the efficiency of the library. In order to

generate an efficient FM library, it may be useful to use a screening means that only allows the growth of clones harbouring a protease gene. A similar methodology was used to isolate *Akkermansia muciniphila* [4] with mucin as a sole carbon source and so it may also be useful to utilise this approach with culture-dependent screens of the gut microbiota to isolate protease producers and potentially novel species. Other carbon sources might include collagen, fibrinogen and Immunoglobulins.

One outstanding factor that this research has generated in chapter 5.0, and chapter 7.0 in particular, stresses the importance of a more direct approach to studying microbial proteases as they are diverse in terms of their catalysis, substrate specificity and regulation. While the method of PCR-based discovery has its drawbacks due to the necessity to have a substantial *a priori* sequence knowledge, if we take a different point of view of aiming to isolate proteases with novel functions instead of striving towards isolating proteases with completely novel sequences, we may be more efficiently able to uncover proteases with clinical and industrial relevance. Purohit *et al*., [5] have demonstrated this process as discussed in chapter 5.0 and in an extensive review describing the process of PCR-based mining and genome walking techniques for isolation of novel bacterial genes by Kotik [6], several studies are highlighted in which genes expressing novel functions were isolated. For example, Hayashi and colleagues [7] utilised a PCR-mining method to isolate novel xylanase genes from the human gut. Uchiyama and Watanabe [8] also alleviated the problem of isolating low copy number genes in environmental DNA samples by developing an inverse PCR using biotin-labelled degenerate primer, affinity purification of the PCR product followed by nested PCR and utilised this techniques to isolate putative chitinase genes, one of which shared less than 35 % nucleotide homology with any other database chitinases.

This could be a useful future direction for studying gut microbial proteases, specifically those that may be significant in terms of disease, as degenerate primers can be designed based on alignments of proteases that are known the act as virulence factors.

5) Aim: To determine optimal conditions for faecal protease extraction and storage over time (Chapter 6.0).

As more and more studies are utilising faecal samples as a means of accessing microbial DNA, metabolites, and protein as well as other molecules, it is becoming increasingly important to assess the parameters that affect the storage of such samples. As the area of interest in this research is predominantly microbial proteases, the effect of long term storage on protein extract protein concentration and protease activity was determined. From the findings of this chapter it can now be advised that for future research of gut microbial proteases, whole faecal samples stored at sub-zero temperatures provide the most robust and repeatable means for accessing microbial protease activity.

6) Aim: To conduct a comprehensive comparison of protease activity in an IBD cohort and compare to a group of healthy volunteers (Chapter 7.0).

One of the most significant findings of this research has been that gut-derived proteases are positively associated with IBD and that they appear to have either an expanded or a different repertoire of functions evidenced by their enhanced and altered substrate specificity. It was also found that different types of proteases were present between the two cohorts all of which was determined by protease inhibitor analysis and specific substrate activity analysis in chapter 7.
Coupling degradomic studies such as in Chapter 7.0 with correct procedure described in chapter 6.0 has helped us to understand the types of proteases that are present in the gut microbiota and how they may be implicated in IBD. From here, projects can now be designed to specifically analyse the groups of proteases to unravel their role in IBD and other diseases. a useful direction for this study is the identification of specific proteases via proteomic analysis analogous to the methods described by Tooth *et al.,* [9] which was used to identify serine proteases in IBS. As this research has discovered that metalloproteases are significantly elevated in an IBD cohort, utilisation of a metal-ligand binding affinity chromatography system to isolate metalloproteases from faecal samples in a similarly designed cohort study would be a significantly useful direction for this research to now take. Specific isolation of types of proteases will facilitate more comprehensive characterisation as well as assisting in proteomic identification and from here, we can fully understand the implications of specific proteases with regards to IBD.

7) Aim: To conduct 16S rRNA community profiling analysis for each cohort (Chapter 7.0)

This research has further validated that there is a dysbiosis of the gut microbiota in IBD and that compositional alterations of the gut microbiota are associated with high levels of protease activity. Specifically, the *Verrucomicrobia* phyla were significantly reduced in the high protease activity cohort and future studies with larger cohorts will help solidify these findings as well as potentially uncovering other significant correlations with other bacterial species such as the members of the *Bacteroidetes* phylum, of which a reduction in the IBD cohort, though not statistically significant, was also observed in this research. While the implications of alterations in the gut microbiota relating to protease activity remain unclear, new hypotheses for future research can be inferred regarding the role of *Verrucomicrobia* in gut health and disease, specifically inflammatory disorders. Research into members of the *Verrucomicrobia* such as *A. muciniphila* and interactions with gut epithelial cells and other bacteria thought to dominate in the absence of *A. muciniphila* will be incredibly useful in unravelling these phenomena and furthering our understanding of the role of the gut microbiota in IBD.

8) Aim: To further unravel the potential role of the gut degradome in the virulence of inflammation (Chapter 7.0)

We have demonstrated that faecal proteases isolated from an IBD cohort are capable of disrupting epithelial barrier function. Repeating this investigation with a larger cohort and conducting experimentation into the specific mechanisms by which these faecal proteases are disrupting barrier inegrity , such as investigating degradation of E-cadherin and other transmembrane proteins and tight junction protein will futher validate this and lead to research into the use of protease inhibitors as for protection of the gut epitheila barrier can be executed. Future research such as this, coupled with research by Motta and colleagues [10] may help futher develop protease inhibtiors as a therapeutic intervention tool for the incidence increasing, debilitating and chronically recurring disorder that is IBD.

8.2 CONCLUDING REMARKS

This research has begun to unravel the molecular mechanisms behind the dysbiosis of the gut microbiota in IBD and this, coupled with future research highlighted in this chapter, could have great implications in the treatment of the disease. Chapter 7.0 demonstrates the significance of microbial proteases with regards to gut health and disease. This emphasises the importance of the research conducted in chapters 3, 4 and 5 in developing tools for isolating and characterising proteases from the gut degradome. This research, and the future research discussed in this chapter, must be pursued if we are to access the full degradome of the gut microbiota; an area of biological research that this thesis has shown is justifiable, valuable and worthwhile.

8.3 REFERENCES

1. Rawlings, N.D., Barrett, A.J., and Bateman, A.(2014). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Research *40*, D343-D350.
2. Macfarlane, G.T., Cummings, J.H., and Allison, C. (1986). Protein Degradation by Human Intestinal Bacteria. Journal of General Microbiology *132*, 1647-1656.
3. Bever, R.A., and Iglewski, B.H. (1988). Molecular characterization and nucleotide sequence of the *Pseudomonas aeruginosa* elastase structural gene. Journal of Bacteriology *170*, 4309-4314.
4. Derrien, M., Vaughan, E.E., Plugge, C.M., and de Vos, W.M. (2004). *Akkermansia muciniphila* gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. International Journal of Systematic and Evolutionary Microbiology *54*, 1469-1476.
5. Purohit, M.K., and Singh, S.P. (2012). A metagenomic alkaline protease from saline habitat: Cloning, over-expression and functional attributes. International Journal of Biological Macromolecules *53*, 138-143.
6. Kotik, M. (2009). Novel genes retrieved from environmental DNA by polymerase chain reaction: Current genome-walking techniques for future metagenome applications. Journal of Biotechnology *144*, 75-82.
7. Hayashi, H., Abe, T., Sakamoto, M., Ohara, H., Ikemura, T., Sakka, K., and Benno, Y. (2005). Direct cloning of genes encoding novel xylanases from the human gut. Canadian Journal of Microbiology *51*, 251-259.
8. Uchiyama, T., and Watanabe, K. (2006). Improved inverse PCR scheme for metagenome walking. BioTechniques *41*, 183-188.
9. Tooth, D., Garsed, K., Singh, G., Marciani, L., Lam, C., Fordham, I., Fields, A., Banwait, R., Lingaya, M., Layfield, R., et al. (2014). Characterisation of faecal protease activity in irritable bowel syndrome with diarrhoea: origin and effect of gut transit. Gut *63*, 753-760.
10. Motta, J.P., Bermudez-Humaran, L.G., Deraison, C., Martin, L., Rolland, C., Rousset, P., Boue, J., Dietrich, G., Chapman, K., Kharrat, P., et al. (2012). Food-grade bacteria expressing elafin protect against inflammation and restore colon homeostasis. Science and Translational Medicine *4*, 158ra144.